TINJAUAN

Aplikasi pembelajaran mesin dalam penemuan dan pengembangan obat

Jessica Vamathevan[1], Dominic Clark[1], Paul Czodrowski [2], Ian Dunham[3], Edgardo Ferran[1], George Lee[4], Bin Li[5], Anant Madabhushi[6,7], Parantu Shah[8], Michaela Spitzer[3] dan Shanrong Zhao[9]*

Abstrak | Jalur penemuan dan pengembangan obat sangat panjang, kompleks, dan bergantung pada banyak faktor. Pendekatan pembelajaran mesin (ML) menyediakan seperangkat alat yang dapat meningkatkan penemuan dan pengambilan keputusan untuk pertanyaan yang ditentukan dengan baik dengan data yang melimpah dan berkualitas tinggi. Peluang untuk menerapkan ML terjadi di semua tahap penemuan obat. Contohnya termasuk validasi target, identifikasi biomarker prognostik, dan analisis data patologi digital dalam uji klinis. Penerapannya beragam dalam konteks dan metodologi, dengan beberapa pendekatan yang menghasilkan prediksi dan wawasan yang akurat. Tantangan penerapan ML terutama terletak pada kurangnya kemampuan interpretasi dan pengulangan hasil yang dihasilkan ML, yang dapat membatasi penerapannya. Di semua bidang, data dimensi tinggi yang sistematis dan komprehensif masih perlu dihasilkan. Dengan upaya berkelanjutan untuk mengatasi masalah ini, serta meningkatkan kesadaran akan faktor-faktor yang diperlukan untuk memvalidasi pendekatan ML, penerapan ML dapat mendorong

1Laboratorium Biologi Molekuler Eropa, Institut Bioinformatika Eropa, Cambridge, Inggris.

2Universitas Teknik Dortmund, Dortmund, Jerman.

3Target Terbuka dan Laboratorium Biologi Molekuler Eropa, Institut Bioinformatika Eropa, Cambridge, Inggris.

4Bristol-Myers Squibb, Princeton, NJ, AS.

5Takeda Pharmaceuticals International Co, Cambridge, MA, USA.

6Case Western Reserve University, Cleveland, OH, AS.

7Louis Stokes Cleveland Veterans Affair Medical Center, Cleveland, OH, AS.

8EMD Serono R&D Institute, Billerica, MA, AS.

9Pfizer Worldwide Research and Development, Cambridge, MA, AS.

**e-mail: jessicav@ebi.ac.uk*

https://doi.org/10.1038/ s41573-019-0024-5

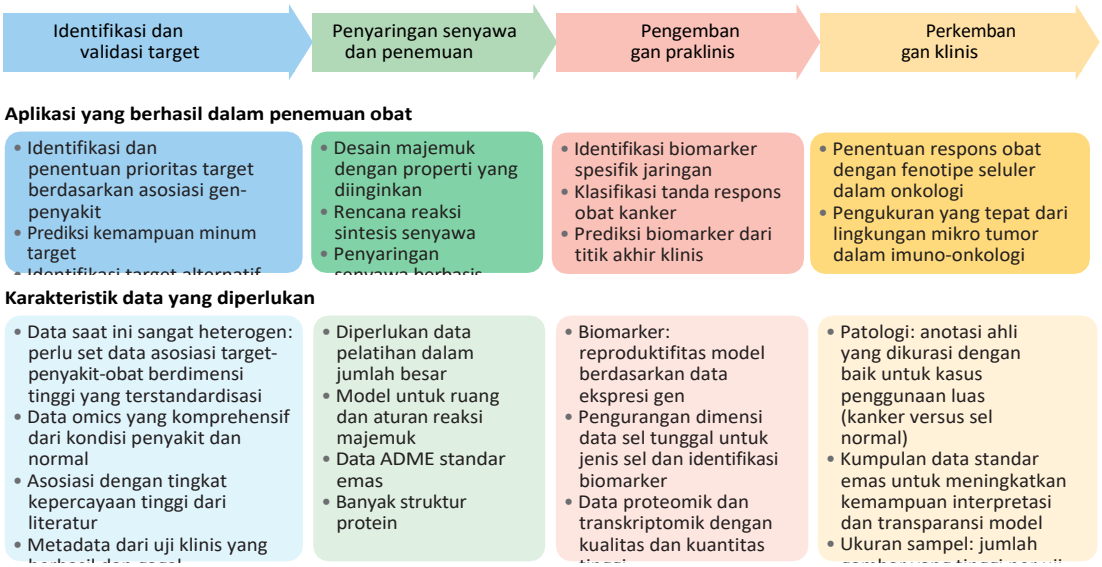
Sistem biologis adalah sumber informasi yang kompleks selama perkembangan dan penyakit. Informasi ini sekarang diukur dan ditambang secara sistematis pada tingkat yang belum pernah terjadi sebelumnya dengan menggunakan sejumlah besar 'omics' dan teknologi pintar. Munculnya pendekatan throughput tinggi terhadap biologi dan penyakit ini menghadirkan tantangan dan peluang bagi industri farmasi, yang bertujuan untuk mengidentifikasi hipotesis terapi yang masuk akal untuk mengembangkan obat. Namun, kemajuan terbaru dalam sejumlah faktor telah meningkatkan minat dalam penggunaan pendekatan pembelajaran mesin (ML) dalam industri farmasi. Ditambah dengan penyimpanan yang dapat diskalakan tanpa batas, peningkatan besar dalam jenis dan ukuran kumpulan data yang dapat menjadi dasar ML telah

memungkinkan perusahaan farmasi untuk mengakses dan mengatur lebih banyak data. Jenis data dapat mencakup gambar, informasi tekstual, biometrik, dan informasi lain dari perangkat yang dapat dikenakan, informasi pengujian, dan data omics berdimensi tinggi1.

Selama beberapa tahun terakhir, bidang kecerdasan buatan (AI) telah beralih dari sebagian besar studi teoritis ke aplikasi dunia nyata. Sebagian besar pertumbuhan eksplosif tersebut berkaitan dengan ketersediaan perangkat keras komputer baru seperti unit pemrosesan grafis (GPU) yang membuat pemrosesan paralel menjadi lebih cepat, terutama dalam komputasi yang intensif secara numerik. Baru-baru ini, kemajuan dalam algoritme ML baru, seperti deep learning (DL)[2], yang membangun Model yang kuat dari data dan keberhasilan yang dapat dibuktikan dari teknik-teknik ini dalam berbagai kontes publik3,4 telah membantu meningkatkan aplikasi ML dalam perusahaan farmasi dalam 2 tahun terakhir.

Meskipun banyak industri layanan konsumen telah menjadi pengadopsi awal metode baru dari bidang ML, penyerapan dari industri farmasi masih tertinggal hingga saat ini. Telah diketahui bahwa tingkat keberhasilan pengembangan obat (sebagaimana didefinisikan dari uji klinis fase I hingga persetujuan obat) sangat rendah di semua area terapeutik dan di seluruh industri farmasi global. Sebuah penelitian terbaru terhadap 21.143 senyawa menemukan bahwa tingkat keberhasilan secara keseluruhan hanya sebesar 6,2%[5]. Oleh karena itu, sebagian besar alasan penggunaan teknologi ML dalam industri farmasi didorong oleh kebutuhan bisnis untuk menurunkan biaya dan gesekan secara keseluruhan.

Semua tahap penemuan dan pengembangan obat, termasuk uji klinis, telah memulai pengembangan dan penggunaan algoritme dan perangkat lunak ML (Gbr. 1) untuk mengidentifikasi target baru[6], memberikan bukti yang lebih kuat untuk asosiasi target-penyakit7, meningkatkan desain dan pengoptimalan senyawa molekul kecil8 , meningkatkan pemahaman tentang mekanisme penyakit, meningkatkan pemahaman tentang fenotipe penyakit dan nonpenyakit9 , mengembangkan penanda biologis baru untuk prognosis, perkembangan, dan kemanjuran obat1 , meningkatkan analisis data biometrik dan data lainnya dari



Gbr. 1 | **Aplikasi pembelajaran mesin dalam jalur penemuan obat dan karakteristik data yang dibutuhkan.** Beberapa aplikasi pembelajaran mesin yang berhasil dalam berbagai tahap pipeline pengembangan obat di perusahaan farmasi telah dipublikasikan. Namun, dalam setiap domain data, masih ada tantangan yang terkait dengan standar kualitas data dan kuantitas data yang diperlukan untuk memanfaatkan potensi penuh dari metode-metode ini untuk penemuan. ADME, penyerapan, distribusi, metabolisme, dan ekskresi.

pemantauan pasien dan perangkat yang dapat dikenakan, meningkatkan pencitraan patologi digital10 dan mengekstrak informasi konten tinggi dari gambar pada semua tingkat resolusi.

Akibatnya, banyak perusahaan farmasi yang mulai berinvestasi dalam sumber daya, teknologi, dan layanan untuk menghasilkan dan mengkurasi kumpulan data untuk mendukung penelitian di bidang ini. Selain itu, perusahaan teknologi raksasa seperti IBM dan Google, perusahaan rintisan bioteknologi, dan pusat-pusat akademis tidak hanya menyediakan layanan komputasi berbasis cloud, tetapi juga bekerja di bidang farmasi dan perawatan kesehatan dengan mitra industri. Ulasan ini memberikan gambaran umum tentang alat dan teknik saat ini (toolbox) yang digunakan dalam ML, termasuk deep neural net, dan gambaran umum tentang kemajuan sejauh ini di bidang aplikasi farmasi utama.

Kotak peralatan pembelajaran mesin

Pada dasarnya, ML adalah praktik menggunakan algoritma untuk mengurai data, mempelajarinya, dan kemudian membuat keputusan atau prediksi tentang keadaan masa depan dari kumpulan data baru. Jadi, alih-alih melakukan rutinitas pengkodean perangkat lunak dengan serangkaian instruksi tertentu (yang telah ditentukan sebelumnya oleh pembuat program) untuk menyelesaikan tugas tertentu, mesin dilatih dengan menggunakan sejumlah besar data dan algoritme yang memberinya kemampuan untuk mempelajari cara melakukan tugas tersebut. Pemrogram mengkodekan algoritme yang digunakan untuk melatih jaringan, bukan mengkodekan aturan ahli.

Algoritme secara adaptif meningkatkan kinerjanya seiring dengan meningkatnya kuantitas dan kualitas data yang tersedia untuk pembelajaran. Oleh karena itu, ML paling baik diterapkan untuk memecahkan masalah yang memiliki sejumlah besar data dan beberapa variabel yang ada, tetapi model atau rumus yang berkaitan

Sedangkan metode unsupervised digunakan untuk tujuan eksplorasi untuk mengembangkan model yang memungkinkan pengelompokan data dengan cara yang tidak ditentukan oleh pengguna. Pembelajaran terawasi melatih model pada hubungan data input dan output yang diketahui sehingga dapat memprediksi output di masa depan untuk input baru. Keluaran di masa depan biasanya berupa model atau hasil untuk klasifikasi data atau pemahaman tentang variabel yang paling berpengaruh (regresi). Teknik pembelajaran tanpa pengawasan mengidentifikasi pola tersembunyi atau struktur intrinsik dalam data input dan menggunakannya untuk mengelompokkan data dengan cara yang bermakna.

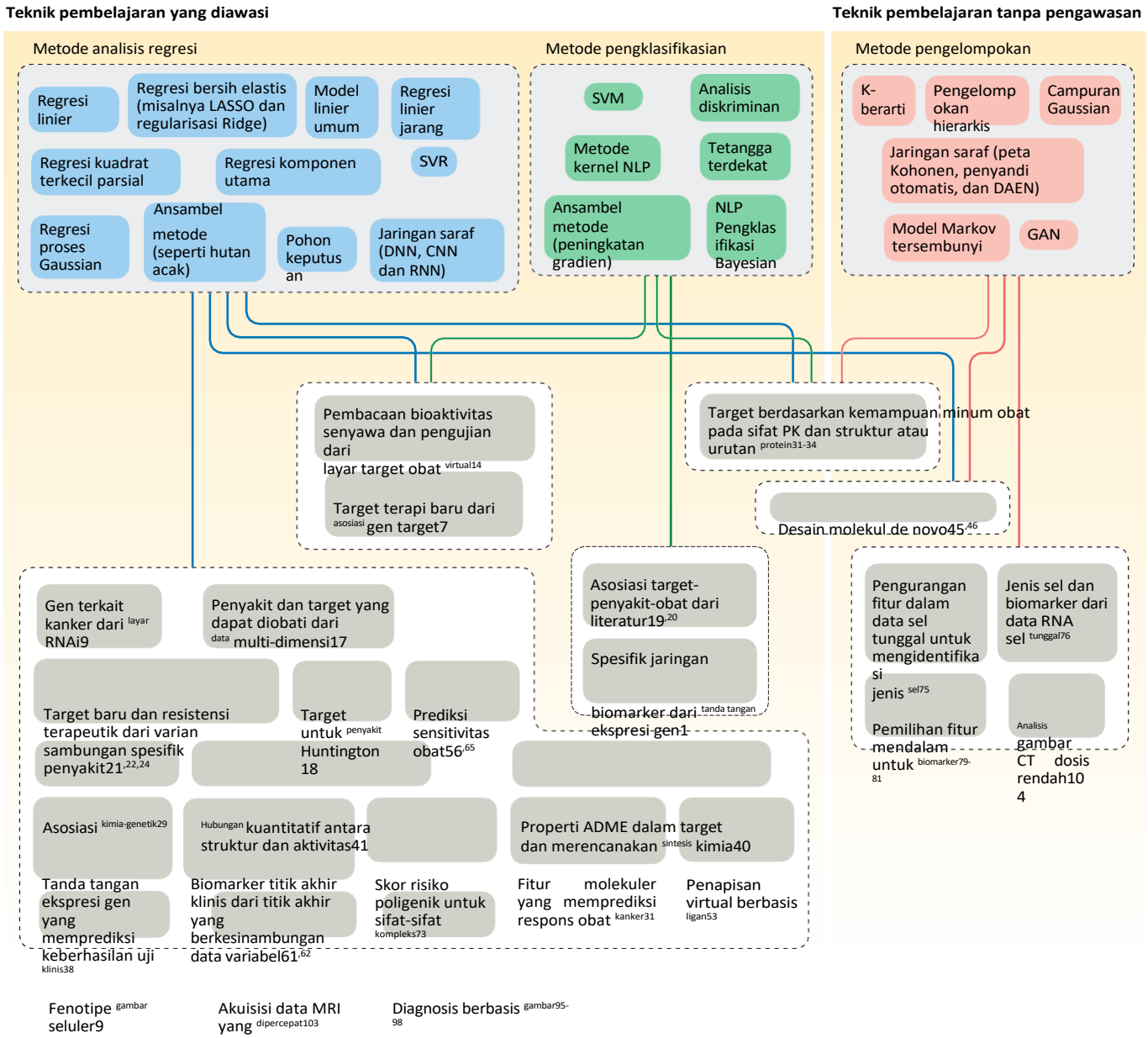
Konsep pemilihan model. Tujuan dari model ML yang baik adalah untuk menggeneralisasi dengan baik dari data pelatihan ke data pengujian yang ada. Generalisasi mengacu pada seberapa baik konsep yang dipelajari oleh model diterapkan pada data yang tidak terlihat oleh model selama pelatihan. Dalam setiap teknik, terdapat beberapa metode (Gbr. 2), yang bervariasi dalam hal akurasi prediksi, kecepatan pelatihan, dan jumlah variabel yang dapat ditangani. Algoritma harus dipilih dengan hati-hati untuk memastikan bahwa algoritma tersebut sesuai dengan masalah yang dihadapi dan jumlah serta jenis data yang tersedia. Jumlah penyetelan parameter yang diperlukan dan seberapa baik metode ini memisahkan sinyal dari derau juga merupakan pertimbangan penting.

Model overfitting terjadi ketika model tidak hanya mempelajari sinyal tetapi juga beberapa fitur yang tidak biasa dari data pelatihan dan memasukkannya ke dalam model, dengan dampak negatif yang dihasilkan pada kinerja model pada data baru. Underfitting mengacu pada model yang tidak dapat memodelkan data pelatihan atau menggeneralisasi data baru. Cara umum untuk membatasi overfitting adalah dengan menerapkan
metode resampling atau untuk menahan sebagian data pelatihan untuk digunakan sebagai kumpulan data validasi. Metode regresi regularisasi (seperti Ridge, LASSO atau jaring elastis) menambahkan penalti pada parameter seiring dengan meningkatnya kompleksitas model sehingga model dipaksa untuk menggeneralisasi data dan tidak

Unit pemrosesan grafis (GPU). Prosesor yang dirancang untuk mempercepat rendering grafis dan dapat menangani puluhan ribu operasi per siklus.

ini tidak diketahui.

Ada dua jenis teknik utama yang digunakan untuk menerapkan ML: pembelajaran terawasi dan tidak terawasi. Metode pembelajaran terawasi digunakan untuk mengembangkan model pelatihan untuk memprediksi nilai masa depan dari kategori data atau



Gbr. 2 | **Alat pembelajaran mesin dan aplikasi penemuan obatnya.** Gambar ini memberikan gambaran umum tentang teknik pembelajaran mesin yang telah digunakan untuk menjawab pertanyaan penemuan obat yang tercakup dalam Tinjauan ini. Berbagai teknik pembelajaran terawasi (metode regresi dan pengklasifikasi) digunakan untuk menjawab pertanyaan yang membutuhkan prediksi kategori data atau variabel kontinu, sedangkan teknik tanpa pengawasan digunakan untuk mengembangkan model yang memungkinkan pengelompokan data. ADME, penyerapan, distribusi, metabolisme, dan ekskresi; CNN, convolutional neural jaringan; CT, tomografi terkomputasi; DAEN, jaringan saraf autoencoder dalam; DNN, jaringan saraf dalam; GAN, jaringan permusuhan generatif; MRI, pencitraan resonansi magnetik; NLP, pemrosesan bahasa alami; PK, farmakokinetik; RNAi, gangguan RNA; RNN, jaringan saraf berulang; SVM, mesin vektor pendukung; SVR, regresi vektor pendukung.

Unit pemrosesan pusat (CPU). Prosesor yang dirancang untuk menyelesaikan setiap masalah komputasi secara umum dan dapat menangani puluhan operasi per siklus. Cache dan memori dirancang agar optimal untuk masalah pemrograman umum apa pun.

Unit pemrosesan tensor (TPU). Co-prosesor yang diproduksi oleh Google yang dirancang untuk mempercepat tugas-tugas pembelajaran mendalam yang dikembangkan menggunakan TensorFlow (kerangka kerja pemrograman) dan dapat menangani hingga 128.000 operasi per siklus.

overfit. Salah satu cara yang paling efektif untuk menghindari overfit adalah metode dropout11, yang secara acak menghapus unit-unit di lapisan tersembunyi. Teknik ML yang berbeda memiliki metrik kinerja yang berbeda. Metrik evaluasi dasar12 seperti akurasi klasifikasi, kappa13 , area di bawah kurva (AUC), logaritmik loss, nilai F1 dan matriks kebingungan dapat digunakan untuk membandingkan performa antar metode. Ketersediaan set data standar emas serta set data yang dibuat secara

independen dapat menjadi sangat berharga dalam menghasilkan model yang berkinerja baik.

Beberapa pustaka perangkat lunak sekarang tersedia untuk komputasi matematika berkinerja tinggi di berbagai platform perangkat keras (central processing unit (CPU)), GPU dan unit pemrosesan tensor (TPU)), dan dari desktop hingga cluster server. Kerangka kerja pro-gramatika ML yang umum digunakan adalah kerangka kerja sumber terbuka TensorFlow, yang awalnya dikembangkan oleh para peneliti dan insinyur dari tim Google Brain di dalam organisasi AI Google (lihat Tautan terkait), serta PyTorch, Keras, dan Scikit-learn.

Arsitektur jaringan saraf dalam. DL adalah reinkarnasi modern dari jaringan saraf tiruan dari tahun 1980-an dan 1990-an dan menggunakan jaringan saraf dalam (DNN) yang canggih dan bertingkat untuk menciptakan sistem yang dapat melakukan deteksi fitur dari sejumlah besar

data pelatihan yang tidak berlabel atau berlabel2. Perbedaan utama antara DL dan jaringan syaraf tiruan tradisional adalah skala dan kompleksitas jaringan yang digunakan. Dalam jaringan saraf, fitur input dimasukkan ke lapisan input, dan setelah sejumlah transformasi nonlinier menggunakan lapisan tersembunyi, prediksi dihasilkan oleh lapisan output. Hal ini biasanya dilakukan dengan menggunakan perambatan balik kesalahan untuk secara progresif mengurangi perbedaan antara nilai output yang diperoleh dan yang diharapkan. Setiap node output berhubungan dengan tugas (atau kelas) yang akan diprediksi. Jika hanya ada satu node di lapisan output, maka jaringan yang sesuai disebut sebagai jaringan saraf tugas tunggal. DL dapat memiliki sejumlah besar lapisan tersembunyi karena menggunakan perangkat keras CPU dan GPU yang lebih kuat, sedangkan jaringan saraf tradisional biasanya menggunakan satu atau dua lapisan tersembunyi karena keterbatasan perangkat keras. Ada juga banyak perbaikan algoritmik dalam DL.

Aplikasi DNN dalam penemuan obat telah banyak dan mencakup prediksi bioaktivitas14, desain molekuler de novo, prediksi sintesis, dan analisis gambar biologis3. Salah satu keuntungan dari DNN adalah bahwa mereka memiliki beberapa arsitektur fleksibel yang berbeda yang dijelaskan di bawah ini dan dengan demikian digunakan untuk menjawab berbagai pertanyaan. Pada arsitektur pertama, deep convolutional neural networks (CNN), beberapa lapisan tersembunyi hanya terhubung secara lokal (bukan global) ke lapisan tersembunyi berikutnya. CNN mencapai kinerja prediktif terbaik di bidang-bidang seperti pengenalan suara dan gambar dengan menyusun fitur-fitur lokal yang sederhana secara hirarkis ke dalam model yang kompleks. Jaringan konvolusi grafik adalah jenis khusus CNN yang dapat diterapkan pada data terstruktur dalam bentuk grafik atau jaringan. Arsitektur kedua adalah jaringan saraf tiruan berulang (RNN), yang mengambil bentuk rantai modul jaringan saraf berulang di mana koneksi antar node membentuk grafik berarah sepanjang urutan. Hal ini memungkinkan analisis perubahan dinamis dari waktu ke waktu di mana informasi yang terus-menerus diperlukan. Jaringan saraf tiruan memori jangka pendek adalah jenis RNN khusus yang mampu mempelajari ketergantungan jangka panjang. Contoh ketiga - jaringan feedforward yang terhubung sepenuhnya - adalah jaringan di mana setiap neuron input terhubung ke setiap neuron di lapisan berikutnya. Ini adalah kebalikan dari RNN di mana, dengan jaringan feedforward yang terhubung sepenuhnya, gradien didefinisikan dengan jelas dan dapat dihitung melalui backpropagation. Model-model ini telah digunakan dalam kasus-kasus pembangunan model prediktif yang menantang, seperti dengan data ekspresi gen, di mana jumlah sampelnya kecil relatif terhadap jumlah fitur. Arsitektur jaringan keempat adalah jaringan saraf deep autoencoder neural network (DAEN). Jenis jaringan saraf ini adalah algoritma pembelajaran tanpa pengawasan yang menerapkan backpropagation untuk memproyeksikan input ke output dengan tujuan pengurangan dimensi15, sehingga mencoba untuk mempertahankan variabel acak yang penting dari data sambil menghilangkan bagian yang tidak penting. Arsitektur jaringan kelima dan terakhir - generative adversarial network (GAN) - terdiri dari dua jaringan (meskipun sering kali merupakan kombinasi dari feedforward neural network dan CNN), di mana satu jaringan ditugaskan untuk menghasilkan konten dan jaringan lainnya untuk mengklasifikasikan konten tersebut.

Karakteristik data. Praktik ML dikatakan terdiri dari setidaknya 80% pemrosesan dan pembersihan data dan 20% aplikasi algoritma. Oleh karena itu, kekuatan prediksi dari setiap pendekatan ML bergantung pada ketersediaan volume data yang tinggi dan berkualitas tinggi. Data yang digunakan untuk pelatihan harus akurat, terkurasi, dan selengkap mungkin untuk memaksimalkan prediktabilitas. Desain eksperimental sering kali melibatkan diskusi tentang ukuran sampel yang ideal dan perhitungan daya yang tepat untuk memperkirakan parameter ini dengan benar. Apakah jenis data yang tepat tersedia dan data apa yang harus dihasilkan secara eksperimental juga merupakan pertimbangan utama untuk pertanyaan-pertanyaan tertentu. Aplikasi ML lebih kuat ketika digunakan pada data yang telah dihasilkan secara sistematik, dengan noise minimal dan anotasi yang baik. Seperti yang akan kita bahas di bawah ini, banyak aplikasi yang tidak terlalu efektif karena data digabungkan dari berbagai sumber dengan kualitas data yang bervariasi. Ada upaya yang sedang berlangsung untuk mengembangkan data beranotasi terbuka di bidang penemuan obat tertentu, seperti [validasi] target16. Hal ini bertujuan untuk menghasilkan anotasi positif dan negatif yang berkualitas baik di bidang yang penting untuk penemuan dan pengembangan obat untuk mendorong penerapan ML.

Aplikasi dalam penemuan obat

Identifikasi dan validasi target. Pendekatan utama dalam penemuan obat adalah mengembangkan obat (molekul kecil, peptida, antibodi, atau modalitas yang lebih baru, termasuk RNA pendek atau terapi sel) yang akan mengubah kondisi penyakit dengan memodulasi aktivitas target molekuler. Terlepas dari kebangkitan baru-baru ini dalam skrining feno-tipik, memulai program pengembangan obat memerlukan identifikasi target dengan hipotesis terapeutik yang masuk akal: bahwa modulasi target akan menghasilkan modulasi keadaan penyakit. Memilih target ini berdasarkan bukti yang tersedia disebut sebagai identifikasi dan penentuan prioritas target. Setelah membuat pilihan awal ini, langkah selanjutnya adalah memvalidasi peran target yang dipilih dalam penyakit menggunakan model ex vivo dan in vivo yang relevan secara fisiologis (validasi target). Meskipun validasi akhir dari target baru akan dilakukan kemudian, melalui uji klinis, validasi target awal sangat penting untuk memfokuskan upaya pada proyek-proyek yang berpotensi berhasil.

Biologi modern semakin kaya akan data. Ini termasuk informasi genetik manusia dalam populasi besar, profil transkriptomik, proteomik, dan metabolomik dari individu yang sehat dan mereka yang memiliki penyakit tertentu serta pencitraan materi klinis dengan konten tinggi. Kemampuan untuk menangkap kumpulan data yang besar ini dan menggunakannya kembali melalui basis data publik menghadirkan peluang baru untuk identifikasi dan validasi target awal. Namun, kumpulan data multi-dimensi ini membutuhkan metode analisis yang tepat untuk menghasilkan model yang valid secara statistik yang dapat membuat prediksi untuk identifikasi target, dan di sinilah ML dapat dieksploitasi. Kisaran eksperimen yang dapat berkontribusi pada identifikasi dan validasi target sangat luas, tetapi jika eksperimen ini digerakkan oleh data, ML semakin banyak diterapkan.

Langkah pertama dalam identifikasi target adalah menetapkan hubungan sebab akibat antara target dan penyakit. Menetapkan hubungan sebab akibat membutuhkan demonstrasi bahwa

pada skala seluruh genom.

Modulasi target mempengaruhi penyakit baik dari variasi yang terjadi secara alamiah (genetik) maupun dari intervensi eksperimental yang dirancang dengan hati-hati. Namun, ML dapat digunakan untuk menganalisis kumpulan data yang besar dengan informasi tentang fungsi dari target yang diduga untuk membuat prediksi tentang kausalitas potensial, yang didorong, misalnya, oleh sifat-sifat dari target yang diketahui. Metode ML telah diterapkan dengan cara ini di beberapa aspek bidang identifikasi target. Costa dkk.[17] membangun meta-classifier berbasis pohon keputusan yang dilatih pada topologi jaringan protein-protein, interaksi metabolik dan transkripsi, serta ekspresi jaringan dan lokalisasi subseluler, untuk memprediksi gen yang terkait dengan morbiditas yang juga dapat diobati. Dengan memeriksa pohon keputusan, mereka mengidentifikasi regulasi oleh beberapa faktor transkripsi (TF), sentralitas dalam jalur metabolisme dan lokalisasi ekstraseluler sebagai parameter kunci. Dalam penelitian lain, model ML berfokus pada penyakit atau area terapi tertentu. Jeon et al.[6] membangun pengklasifikasi mesin vektor pendukung (SVM) menggunakan berbagai set data genom untuk mengklasifikasikan protein ke dalam target obat dan target non-obat untuk kanker payudara, pankreas, dan ovarium. Fitur klasifikasi utama adalah esensialitas gen, ekspresi mRNA, jumlah salinan DNA, kejadian mutasi dan topologi jaringan interaksi protein-protein. Secara keseluruhan, 122 target kanker global diidentifikasi, 69 di antaranya tumpang tindih dengan 116 target kanker yang telah diketahui. Selain itu, 266, 462 dan 355 target diidentifikasi sebagai spesifik untuk kanker payudara, pankreas dan ovarium. Dua target yang diprediksi telah divalidasi dengan inhibitor peptida yang memiliki efek anti-proliferasi yang kuat dalam model kultur sel. Lebih lanjut, inhibitor untuk 137 target kanker pankreas yang diprediksi hampir dua kali lebih mungkin menunjukkan penghambatan yang kuat terhadap kelangsungan hidup sel dibandingkan senyawa lainnya. Ament et al.[18] membangun model berdasarkan situs pengikatan TF tikus dan data profil transkriptom untuk mengkarakterisasi perubahan transkripsi yang mendasari penyakit Huntington. Mereka merekonstruksi model skala genom dari gen target untuk 718 TF di striatum tikus menggunakan model regresi dan regularisasi LASSO. Secara keseluruhan, 13 dari 48 modul TF yang diidentifikasi diekspresikan secara berbeda dalam jaringan striatal pada penyakit manusia dan memberikan titik awal yang potensial untuk terapi Huntington. Target molekuler untuk terapi anti-penuaan spesifik jaringan telah diidentifikasi oleh Mamoshina et al.[1]. Mereka membandingkan tanda tangan ekspresi gen dari otot muda dan otot tua. Perbandingan beberapa metode ML yang telah dilihat menunjukkan bahwa SVM dengan kernel linier dan seleksi fitur mendalam paling cocok untuk identifikasi biomarker penuaan. Dalam setiap contoh ini, ML menghasilkan serangkaian prediksi target yang memiliki sifat yang menunjukkan bahwa mereka cenderung mengikat obat, atau terlibat dalam penyakit, tetapi validasi lebih lanjut sangat penting untuk menghasilkan hipotesis terapeutik.

Literatur adalah sumber utama pengetahuan tentang asosiasi target dengan penyakit. Pemrosesan literatur secara otomatis membuka informasi dari teks yang tidak terstruktur yang seharusnya tidak dapat diakses.

Pengklasifikasi mesin vektor pendukung (SVM)
Metode yang melakukan tugas klasifikasi dengan membuat garis pemisah untuk membedakan antara objek dengan keanggotaan kelas yang berbeda dalam ruang multi dimensi.

CLIP-seq
Ultraviolet crosslinking immunoprecipitation (CLIP) diikuti dengan pengurutan RNA untuk mengidentifikasi semua spesies RNA yang terikat oleh protein yang diinginkan. Metode ini dapat digunakan untuk memetakan situs pengikatan protein RNA atau situs modifikasi RNA

Kemajuan terbaru dalam pemrosesan bahasa alami (NLP), sebuah pendekatan ML yang diterapkan pada penambangan teks, telah memungkinkan penambangan data yang lebih efektif untuk mengidentifikasi makalah yang relevan. BeFree19 menerapkan metode NLP Kernel untuk mengidentifikasi asosiasi obat-penyakit, gen-penyakit, dan target-obat di Medline abstrak. Pendekatan pembelajaran yang diawasi ini bergantung pada korpus basis data reaksi obat yang merugikan Uni Eropa (EU-ADR) yang dianotasi secara manual dan korpus yang dianotasi secara semi-otomatis berdasarkan Basis Data Asosiasi Genetik. DigSee20 mengidentifikasi gen dan penyakit dalam abstrak Medline, menggunakan NLP untuk mengekstrak kejadian biologis di antara entitas-entitas ini dan mengurutkan kalimat-kalimat bukti dengan pengklasifikasi Bayesian.

Salah satu area dengan cakupan yang luas untuk ML adalah dalam memahami aspek-aspek dasar biologi untuk mengidentifikasi peluang terapeutik melalui modalitas alternatif atau target baru. Memahami variasi genetik dalam sinyal penyambungan adalah salah satu contohnya. Model penyambungan DL sekarang dapat secara akurat memprediksi sinyal penyambungan alternatif21. Model penyambungan inte- gratif terbaru22 menggabungkan data uji CLIP-seq pengikatan faktor penyambungan secara in vivo dengan eksperimen pengurutan RNA di mana faktor penyambungan ini telah dirobohkan atau diekspresikan secara berlebihan. Menggabungkan model kode penyambungan dengan prediksi variasi penyambungan de novo dan kompleks telah memungkinkan identifikasi varian penyambungan yang spesifik untuk penyakit Alzheimer23. Aplikasi terbaru dari pendekatan serupa mengidentifikasi mekanisme pelarian dari imunoterapi CART-1924, varian genetik langka yang menyebabkan ketulian25 dan varian penyambungan yang terkait dengan autisme26.

ML juga dapat memprediksi efek obat spesifik kanker. Iorio dkk.[27] menyaring 990 garis sel kanker terhadap 265 obat antikanker dan menyelidiki bagaimana ekspresi gen di seluruh genom, metilasi DNA, jumlah salinan gen, dan data mutasi somatik memengaruhi respons obat. Mereka menggunakan ANOVA, model logika dan algoritme ML (regresi jaring elastis dan hutan acak) untuk mengidentifikasi fitur molekuler yang memprediksi respons obat. Jenis data yang paling prediktif di seluruh jenis kanker adalah ekspresi gen, sedangkan model spesifik kanker yang paling prediktif mencakup fitur genomik (mutasi penggerak atau perubahan jumlah salinan) dan bahkan lebih baik lagi jika menyertakan data metilasi DNA. Tsherniak et al.[28] menggunakan data dari skrining interferensi RNA (RNAi) dari 501 garis sel kanker untuk menemukan penanda molekuler yang mendahului ketergantungan kanker terhadap 769 gen. Mereka mengembangkan model regresi nonlinier berdasarkan pohon inferensi bersyarat untuk menghasilkan model prediktif berdasarkan ekspresi gen, jumlah salinan gen, dan mutasi gen somatik. McMillan dkk.[29] menyaring 222 bahan kimia terhadap
>100 model sel yang sangat beranotasi dari lesi kanker paru somatik yang beragam dan khas. Mereka menerapkan ML yang teratur (jaring elastis) dan metrik berbasis probabilitas (pemindaian Kolmogorov-Smirnov) untuk mengidentifikasi 171 asosiasi genetik-kimiawi yang mengungkapkan kerentanan mekanistik yang dapat ditargetkan pada berbagai onkotipe yang tidak dapat diobati dengan terapi yang efektif. Pendekatan ini menunjukkan bahwa ada peluang untuk pengobatan presisi intrinsik tumor.

Pertanyaan penting lainnya bagi para pengembang obat adalah seberapa besar kemungkinan obat tersebut dapat dibuat untuk target tertentu. Untuk obat molekul kecil, hal ini memerlukan identifikasi target yang memiliki fitur yang menunjukkan bahwa protein ini dapat mengikat molekul kecil30. Atribut target yang berbeda dapat digunakan untuk menghasilkan model kemampuan obat ini. Nayal dan Honig31 melatih pengklasifikasi hutan acak pada atribut fisikokimia, struktural dan geometris dari 99 pengikatan obat

dan 1.187 rongga pengikat non-obat dari satu set 99 protein. Ukuran dan bentuk rongga permukaan adalah fitur yang paling penting. Beberapa penelitian memperoleh berbagai sifat fisikokimia dari sekuens protein dari target obat dan non-obat yang diketahui dan menerapkan SVM[32,33] atau SVM bias dengan autoencoder bertumpuk, model DL[34], untuk memprediksi target yang dapat diberi obat. Protein yang dapat diminum juga telah ditemukan untuk menempati wilayah tertentu dari jaringan interaksi protein-protein dan cenderung sangat terhubung[6,17,35]. Sekali lagi, contoh-contoh pendekatan ML ini menghasilkan kumpulan target yang diprediksi kemungkinan besar akan mengikat obat, sehingga mengurangi ruang pencarian potensial, tetapi target-target ini memerlukan validasi lebih lanjut.

Cawan suci untuk identifikasi atau validasi target adalah prediksi awal keberhasilan uji klinis di masa depan untuk program penemuan obat berbasis target. Berbagai analisis non-ML menunjukkan kemungkinan prediktor keberhasilan[5,36,37]. Dengan menggunakan ML, Rouillard et al.[38] menilai data omics untuk satu set 332 target yang berhasil atau gagal dalam uji klinis fase III dengan pemilihan fitur multivariat. Mereka menemukan data ekspresi gen secara khusus dapat memprediksi target yang berhasil, ditandai dengan ekspresi RNA rata-rata yang rendah dan varians yang tinggi di seluruh jaringan. Studi ini mengkonfirmasi temuan sebelumnya bahwa target yang ideal menunjukkan ekspresi spesifik penyakit pada jaringan yang terkena[39]. Ferrero et al.[7] melatih berbagai pengklasifikasi ML menggunakan asosiasi target-penyakit dari platform target terbuka16 untuk memprediksi target terapeutik potensial de novo. Penilaian pentingnya fitur mengidentifikasi keberadaan model hewan, ekspresi gen dan data genetik sebagai jenis data utama untuk prediksi target terapeutik yang tidak tergantung pada indikasi. Namun, pendekatan ini dibatasi oleh sifat data yang jarang dan kurangnya informasi tentang alasan kegagalan program yang dimulai. Lebih mendasar lagi, karena lamanya waktu antara memulai program penemuan obat yang sukses dan membawa obat tersebut ke pasar, program yang sukses mencerminkan paradigma sebelumnya untuk pengembangan obat. Pendorong keberhasilan program molekul kecil tidak mungkin sama saat ini, karena modalitas yang lebih baru, seperti biologis (termasuk antibodi), telah tersedia. Meningkatnya fokus pada pengobatan presisi memperkenalkan kendala tambahan. Sangat penting untuk pendekatan prediksi masa depan bahwa data yang luas tentang program penemuan obat yang berhasil dan gagal tersedia dengan metadata dalam domain publik.

Desain dan optimasi molekul kecil. *Penemuan* kandidat obat yang dapat memblokir atau mengaktifkan protein target yang diminati melibatkan penyaringan pustaka senyawa besar secara virtual dan eksperimental dengan hasil tinggi yang ekstensif. Struktur kandidat kemudian disempurnakan dan dimodifikasi lebih lanjut untuk meningkatkan spesifisitas dan selektivitas target, bersama dengan sifat farmako-dinamis, farmakokinetik, dan toksikologi yang dioptimalkan. Namun, yang terpenting, kurangnya data berkualitas tinggi yang memadai untuk kimia baru seperti proteolysis-targeting chi- meras (PROTAC) dan makrosiklus dapat membatasi dampak ML pada kimia tersebut.

Banyak penelitian telah dilakukan untuk menerapkan metode DL, seperti jaringan saraf tiruan multi-tugas, pada penyaringan virtual berbasis ligan. Diberikan senyawa timbal, senyawa yang memiliki struktur kimia yang serupa dapat diidentifikasi

Metode heuristik
Fungsi yang menghitung perkiraan biaya dari suatu masalah (atau mengurutkan alternatif.)

secara komputasi. Hal ini biasanya dilakukan dengan menggunakan metode statistik klasik, tetapi DNN multi-tugas terbukti lebih efektif[40]. DNN dapat secara signifikan meningkatkan daya prediksi ketika menyimpulkan sifat dan aktivitas molekul kecil[41]. Teknik pembelajaran sekali tembak dapat digunakan untuk secara substansial mengurangi jumlah data yang diperlukan untuk membuat prediksi yang berarti tentang pembacaan molekul dalam pengaturan eksperimental baru. Menggabungkan ML dengan model keadaan Markov, teknik ini digunakan untuk mengidentifikasi mekanisme pengikatan opiat yang sebelumnya tidak diketahui pada reseptor μ-opioid, yang mengungkapkan situs alosterik yang terlibat dalam aktivasi[42]. Namun, manfaat model multi-tugas dibandingkan model tugas tunggal sangat bergantung pada kumpulan data. Untuk membantu membandingkan algoritma ML, Pande dkk. menyusun kumpulan data pembanding yang besar, MoleculeNet[43], yang telah digunakan untuk perbandingan algoritma ML yang berbeda. MoleculeNet berisi data tentang sifat-sifat lebih dari 700.000 senyawa. Semua kumpulan data telah dikurasi dan diintegrasikan ke dalam paket sumber terbuka DeepChem (lihat Tautan terkait), yang juga menyertakan alat lain.

DNN dan algoritme pencarian pohon modern juga dapat digunakan untuk merencanakan rute sintesis kimia yang efisien. Untuk merencanakan sintesis molekul target, molekul secara formal diuraikan menggunakan reaksi terbalik (retrosintesis). Prosedur ini menghasilkan urutan reaksi yang kemudian dapat dijalankan di laboratorium dalam arah maju untuk mensintesis target. Tantangan utamanya adalah menerapkan pengetahuan kimia sintetis secara sistematis pada proses ini. Penggabungan aturan transformasi secara manual menjadi penghalang karena pengetahuan kimia tumbuh secara eksponensial, dan ruang lingkup serta keterbatasan banyak reaksi tidak sepenuhnya dipahami. Untuk mengekstrak aturan secara otomatis, Segler dkk.[44] menggunakan database Reaxys (~11 juta reaksi dan ~300.000 aturan) dan melakukan pencarian pohon Monte Carlo (MCTS) untuk menilai simpul-simpul pohon bersama dengan DNN untuk mengarahkan pencarian ke arah yang paling menjanjikan. Dalam analisis kuantitatif, metode ini mengungguli standar emas, pencarian pertama yang terbaik, dengan dua implementasi yang berbeda (metode heuristik dan neural). Selain itu, MCTS 30 kali lebih cepat daripada metode pencarian berbantuan komputer tradisional untuk hampir dua pertiga molekul yang diperiksa. Tes kualitatif juga dilakukan dalam studi double-blind. Ahli kimia organik diminta untuk memilih antara rute sintesis berbasis literatur dan prediksi tanpa mengetahui bagaimana rute tersebut diperoleh. Di sini, untuk pertama kalinya, para ahli kimia menganggap kualitas rute yang diprediksi rata-rata sama baiknya dengan rute yang diambil dari literatur.

Aplikasi DL lainnya yang berharga adalah desain molekuler de novo melalui pembelajaran penguatan. Para peneliti di AstraZeneca[45] memanfaatkan RNN untuk perluasan ruang kimia dengan menyetel model generatif berbasis urutan untuk mendesain senyawa dengan nilai yang hampir optimal untuk kelarutan, kesesuaian farmakokinetik, bioaktivitas, dan parameter lainnya. Kadurin dkk.[46] juga mengembangkan model serupa menggunakan deep GANs untuk melakukan ekstraksi fitur molekuler pada set data yang sangat besar. Namun, perlu dicatat bahwa pembelajaran penguatan mungkin tidak membantu dalam mengidentifikasi rute sintetis yang baru dan belum pernah ada sebelumnya47.

Kompetisi pemecahan masalah komunitas dapat berguna untuk memajukan pengembangan metode di bidang tertentu. Para peneliti di Merck Sharp & Dohme mensponsori kompetisi Kaggle untuk prediksi parameter penyerapan, distribusi, metabolisme, dan ekskresi (ADME) yang berkaitan serta beberapa target biokimia. Tim pemenang menggunakan DNN, yang, dalam 13 dari 15 sistem pengujian, berkinerja sedikit lebih baik daripada hutan acak standar[41]. Beberapa pembelajaran utama mereka adalah bahwa optimalisasi hyperparameter dapat meningkatkan DNN, pemilihan fitur tidak diperlukan, model multi-tugas per-bentuk lebih baik daripada model tugas tunggal dan overfitting dapat dicegah dengan menggunakan dropout. Ramsundar dkk.[40] juga mengamati bahwa DNN multi-tugas berkinerja lebih baik daripada DNN tugas tunggal. Perbandingan antara DNN tugas tunggal dan multi-tugas serta perbandingan antara metode ML yang berbeda (hutan acak, SVM, naif Bayes, dan regresi logika) dilakukan oleh Lenselink dkk.[48] menggunakan satu set data standar yang diperoleh dari ChEMBL[49]. Di sini, model DNN memiliki kinerja terbaik, dan DNN multi-tugas juga ditemukan lebih baik daripada DNN tugas tunggal. DNN multi-tugas juga terbukti lebih baik untuk prediksi optimasi timbal dan identifikasi timbal, karena mereka dapat mensintesis informasi dari berbagai sumber biologis yang berbeda[50] karena adanya beberapa node di lapisan keluaran.

Pemilihan fitur sebelum membangun model dapat meningkatkan model ML, seperti yang ditunjukkan dalam sebuah studi oleh Kramer dan Gütlein[51]. Mereka juga dapat mendeteksi peningkatan dalam model hutan acak terhadap metode ML lainnya seperti SVM dan naive Bayes, dengan kinerja yang lebih cepat dan lebih sedikit fitur yang digunakan saat melatih model. Menurut mereka, salah satu manfaat utama dari penyaringan bit sidik jari kimiawi adalah peningkatan kemampuan interpretasi model. Jika sidik jari tidak disaring, kemampuan interpretasi akan terhambat karena efek yang disebut 'tabrakan bit'. Dampak penting dari penyaringan sidik jari juga ditunjukkan secara independen oleh Landrum dkk.[8].

Hochreiter et al.[52] juga menemukan bahwa mod berbasis DNN secara signifikan mengungguli semua metode yang bersaing dan bahwa kinerja prediktif DL, menggunakan kumpulan data dari semua tes ChEMBL dan prediksi target berdasarkan input sistem entri jalur input molekuler yang disederhanakan (SMILES) input, dalam banyak kasus sebanding dengan tes yang dilakukan di laboratorium basah. Kelompok Hochreiter juga menunjukkan bahwa DNN mengungguli semua metode ML lainnya (k-tetangga terdekat, naif Bayes, hutan acak dan SVM) dan metode berbasis statistik (similarity ensem-pendekatan ble[53]) untuk prediksi target[54]. Kelompok yang sama memenangkan sebagian besar tantangan dalam Tox21 Data Challenge 2014 (REF.[55]).

Biomarker prediktif. Penemuan biomarker berbasis ML dan model prediktif sensitivitas obat merupakan pendekatan yang telah ditunjukkan untuk membantu meningkatkan tingkat keberhasilan klinis, untuk lebih memahami mekanisme kerja obat dan untuk mengidentifikasi obat yang tepat untuk pasien yang tepat[56-58]. Uji klinis tahap akhir membutuhkan waktu bertahun-tahun dan jutaan dolar untuk dilakukan, sehingga akan sangat bermanfaat untuk membangun, memvalidasi, dan menerapkan model prediktif lebih awal, dengan menggunakan data uji klinis praklinis dan / atau tahap awal. Biomarker translasi dapat diprediksi dengan menggunakan pendekatan ML pada set data praklinis. Setelah divalidasi menggunakan set data independen (baik praklinis maupun klinis), model dan biomarker yang sesuai dapat diterapkan untuk mengelompokkan pasien, mengidentifikasi indikasi potensial, dan menyarankan mekanisme kerja obat (Gbr. 4). Meskipun ada ribuan makalah tentang biomarker dan model prediktif dalam literatur, hanya sedikit yang telah digunakan dalam uji klinis. Berbagai faktor berkontribusi terhadap kesenjangan ini, termasuk kualitas data, pemilihan model, akses ke data dan perangkat lunak, reproduktifitas model, dan desain pengujian yang sesuai untuk pengaturan klinis. Untuk mengatasi beberapa masalah terkait model, beberapa upaya komunitas telah mengevaluasi pendekatan ML untuk mengembangkan model klasifikasi dan regresi. Beberapa tahun yang lalu, Badan Pengawas Obat dan Makanan AS (FDA) menyelenggarakan inisiatif MicroArray Quality Control II (MAQC II) untuk mengevaluasi berbagai metode ML untuk memprediksi titik akhir klinis dari data ekspresi gen awal[59]. Dalam proyek ini, 36 tim independen menganalisis 6 set data microarray untuk menghasilkan model prediktif untuk mengklasifikasikan sampel dengan 1 dari 13 titik akhir klinis. Pengamatan umum termasuk pentingnya proses kontrol kualitas data, kebutuhan akan ilmuwan yang terampil (beberapa tim berkinerja lebih baik secara konsisten daripada tim lain yang menggunakan metode ML yang sama) dan pentingnya memilih pendekatan pemodelan yang tepat untuk titik akhir klinis. Sebagai contoh, prediksi kelangsungan hidup yang buruk secara keseluruhan untuk pasien dengan multiple myeloma bisa jadi sebagian disebabkan oleh penerapan batas waktu kelangsungan hidup yang sewenang-wenang selama 24 bulan. Baik ekspresi gen maupun kelangsungan hidup secara keseluruhan pada multiple myeloma merupakan variabel kontinu, dan oleh karena itu, model prediksi berbasis regresi sangat sesuai. Memang, dengan menggunakan pendekatan regresi Cox univariat, tanda ekspresi gen yang secara signifikan memprediksi subkelompok pasien berisiko tinggi telah diidentifikasi[60]. Tanda tangan ini dikonfirmasi dalam beberapa penelitian independen dan dari pendekatan berbasis regresi yang berbeda[61-64],

Tantangan yang belum terselesaikan dalam bidang desain molekul kecil adalah bagaimana cara terbaik untuk merepresentasikan struktur kimia. Banyak sekali representasi yang ada, mulai dari sidik jari lingkaran

Sidik jari kimia
Sebuah konsep yang digunakan dalam informatika kimia untuk membandingkan molekul satu sama lain. Struktur molekul dikodekan dalam serangkaian digit biner (bit) yang merepresentasikan ada atau tidaknya substruktur tertentu dalam molekul.

Sistem entri baris masukan molekul y a n g disederhanakan (SMILES) Notasi baris untuk memasukkan dan merepresentasikan molekul
dan reaksi; misalnya, karbon dioksida direpresentasikan sebagai O= C= O.

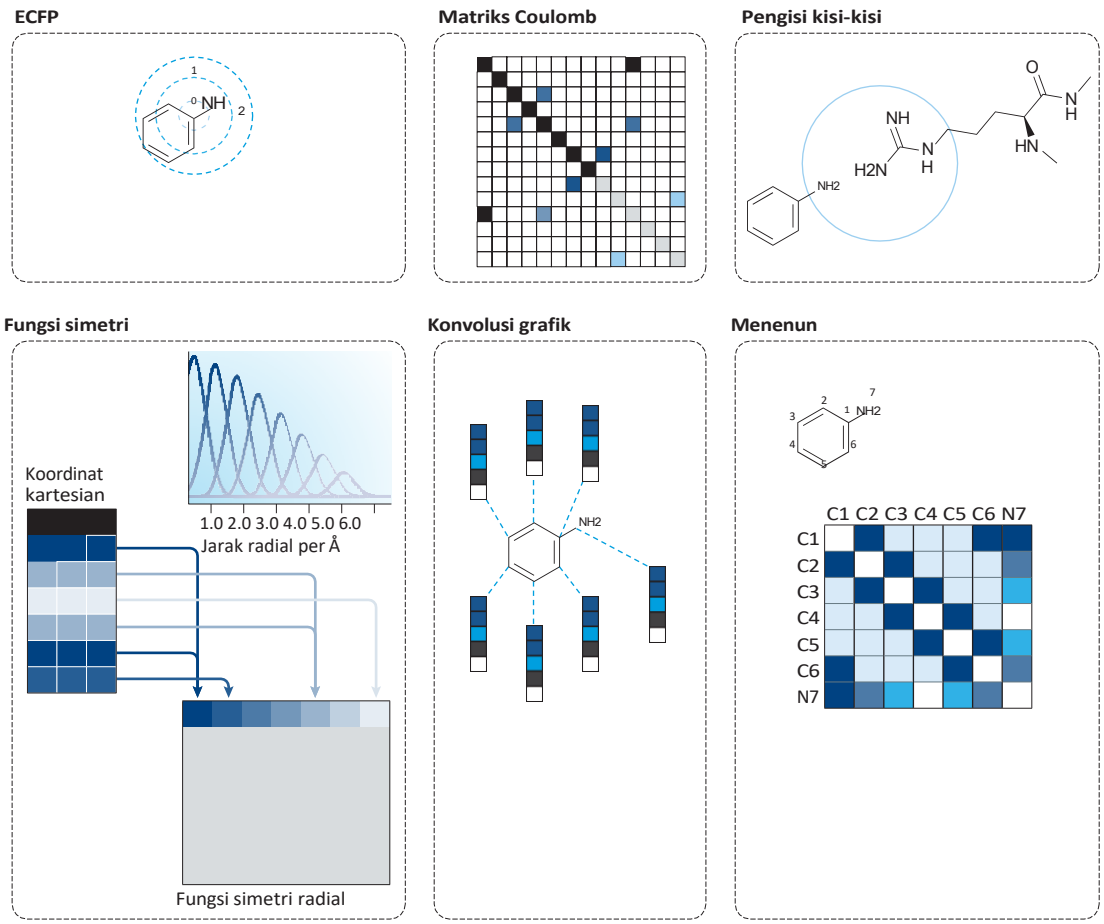
sederhana seperti sidik jari konektivitas yang diperluas (ECFP) hingga fungsi simetri yang canggih (Gbr. 3). Masih belum jelas representasi struktur mana yang paling cocok untuk masalah desain molekul kecil. Oleh karena itu, akan menarik untuk melihat apakah peningkatan studi ML di bidang cheminformatics akan memberikan lebih banyak panduan tentang pilihan terbaik untuk representasi struktur.

menyoroti keuntungan dari pendekatan regresi tanpa keanggotaan kelas yang telah ditentukan sebelumnya.

Penelitian National Cancer Institute (NCI)-DREAM merupakan upaya komunitas lainnya untuk mengevaluasi metode regresi untuk membangun model prediksi sensitivitas obat (yang didefinisikan sebagai pertanyaan regresi)[65] . Setiap tim yang berpartisipasi menggunakan pendekatan pemodelan terbaik mereka dan mengoptimalkan set parameter mereka pada set data pelatihan yang sama (35 garis sel kanker payudara yang diobati dengan 31 obat) kemudian menguji kinerja model mereka pada set data pengujian yang sama (18 garis sel kanker payudara yang diobati dengan 31 obat yang sama). Enam jenis data profil dasar tersedia untuk menghasilkan model prediktif - microarray RNA, larik polimorfisme nukleotida tunggal (SNP), pengurutan RNA, fase terbalik

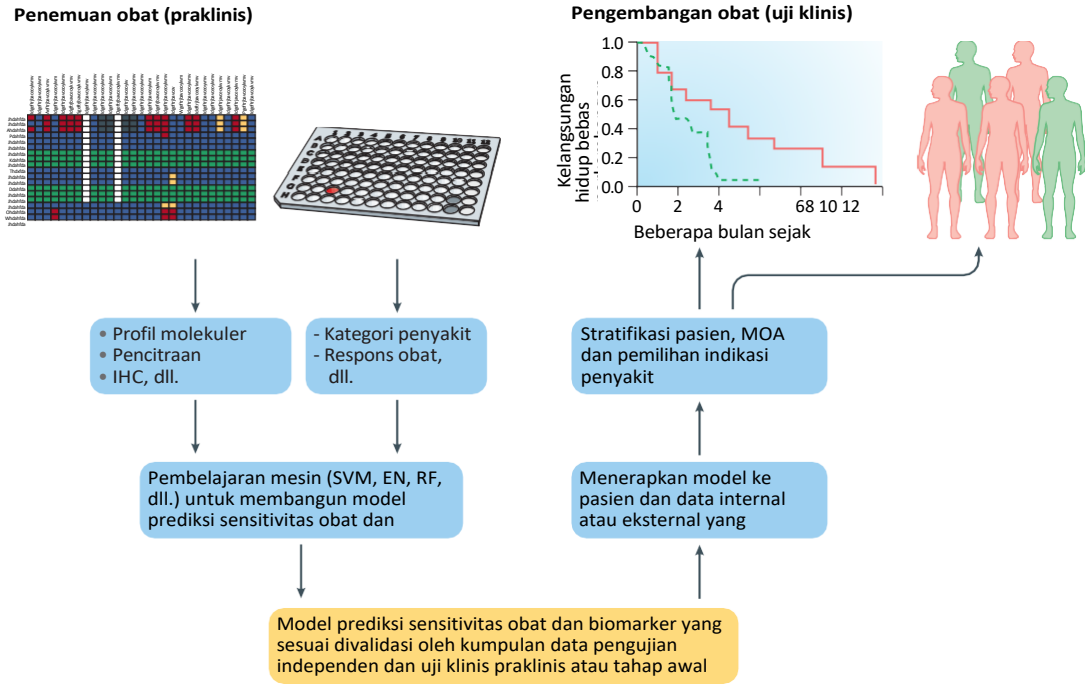
array protein, sekuensing eksom dan status metilasi DNA - di mana 44 tim yang berpartisipasi menggunakan berbagai pendekatan regresi seperti metode kernel, regresi nonlinier (pohon regresi), regresi linier jarang, regresi kuadrat terkecil parsial, regresi komponen utama, atau metode ansambel. Konsisten dengan hasil MAQC II, beberapa tim secara konsisten mengungguli tim lain yang menggunakan pendekatan yang sama. Kinerja yang berbeda kemungkinan mencerminkan rincian teknis yang digunakan untuk kontrol kualitas, reduksi data, pemilihan fitur, strategi pemisahan dan penyempurnaan parameter ML, serta potensi penggabungan pengetahuan biologis seperti informasi fungsi gen atau data klinis ke dalam konstruksi model prediktif. Selain itu, beberapa obat lebih mudah untuk membangun model prediktif daripada yang lain untuk semua tim dan metode. Kumpulan data dan hasil tantangan NCI-DREAM terus berlanjut untuk digunakan sebagai set data validasi untuk pengembangan dan evaluasi metode, misalnya, pada kerangka kerja random for- est ensemble66 , analisis faktor kelompok67 dan pendekatan lainnya68,69.

Beberapa studi kasus yang sukses kini telah dipublikasikan di mana model prediktif yang dihasilkan ML dan biomarker yang sesuai telah memainkan peran penting dalam penemuan dan pengembangan obat. Li et al.[56] melakukan studi kasus menggunakan obat standar perawatan di mana mereka pertama kali membangun model untuk sensitivitas obat terhadap erlotinib dan sorafenib (satu model untuk setiap obat) menggunakan data skrining garis sel kanker. Mereka kemudian menerapkan model tersebut untuk mengelompokkan pasien dari uji klinis BATTLE70 , yang diobati dengan salah satu dari dua obat tersebut, dan menunjukkan bahwa model tersebut bersifat prediktif dan spesifik terhadap obat. Gen biomarker yang diturunkan dari model terbukti mencerminkan mekanisme



Gbr. 3 | **Tantangan representasi struktur senyawa dalam model pembelajaran mesin.** Representasi yang tepat dari struktur kimia dan fitur-fiturnya dapat mengambil banyak representasi tergantung pada aplikasi yang diperlukan. Sidik jari konektivitas yang diperluas (ECFP) berisi informasi tentang karakteristik topologi molekul, yang memungkinkan informasi ini diterapkan pada tugas-tugas seperti pencarian kemiripan dan prediksi aktivitas.
Matriks Coulomb mengkodekan informasi tentang muatan nuklir molekul dan koordinatnya. Metode fitur kisi-kisi menggabungkan fitur struktural ligan dan protein target serta gaya antarmolekul yang berkontribusi pada afinitas pengikatan. Fungsi simetri adalah pengkodean umum lain dari informasi koordinat atom, yang berfokus pada jarak antara pasangan atom dan sudut yang terbentuk dalam kembar tiga atom. Metode konvolusi grafik menghitung vektor fitur awal dan daftar tetangga untuk setiap atom yang merangkum lingkungan kimiawi lokal atom, termasuk jenis atom, jenis hibridisasi, dan struktur valensi. Fiturisasi menenun menghitung vektor fitur untuk setiap pasangan atom dalam molekul, termasuk sifat ikatan (jika terhubung langsung), jarak grafik dan info cincin, membentuk matriks fitur.

Direproduksi dengan izin dari Royal Society of Chemistry, Wu, Z. et al.
MoleculeNet: tolok ukur untuk pembelajaran mesin molekuler. *Chem. Sci.* **9**, 513-530 (2018), REF.[43].



Gbr. 4 | **Memanfaatkan biomarker prediktif untuk mendukung penemuan dan pengembangan obat.** Model prediksi sensitivitas obat (kotak kuning) dapat dibuat dengan menggunakan pendekatan pembelajaran mesin pada data praklinis. Model tersebut kemudian dapat diuji menggunakan data dari sampel pasien klinis tahap awal. Setelah divalidasi, model tersebut dapat digunakan untuk stratifikasi pasien dan/atau pemilihan indikasi penyakit untuk mendukung pengembangan klinis obat, serta untuk menyimpulkan mekanisme kerjanya. EN, jaring elastis; IHC, imunohistokimia; MOA, mekanisme aksi; RF, hutan acak; SVM, mesin vektor pendukung.

Ketika dikombinasikan dengan data domain publik yang dinormalisasi secara global dari berbagai jenis kanker, model ini memprediksi sensitivitas jenis kanker terhadap setiap obat yang konsisten dengan indikasi yang disetujui FDA. Studi ini menunjukkan bahwa menggunakan pendekatan ML untuk mengidentifikasi fitur-fitur utama yang berkontribusi terhadap sensitivitas obat di berbagai jenis kanker dengan cara yang tidak bergantung pada jaringan dapat berguna untuk pengembangan obat (dibandingkan dengan uji klinis berbasis jenis kanker yang diikuti dengan perluasan label). Pada tahun 2017, FDA menyetujui pembrolizumab penghambat kematian sel terprogram 1 (PD1) untuk kanker dengan biomarker genetik tertentu. Ini adalah persetujuan FDA pertama yang didasarkan pada biomarker genetik indikasi silang dan bukan pada jenis kanker[71], yang menyoroti perlunya lebih banyak penemuan biomarker berbasis mekanisme.

Baru-baru ini, ada banyak kemajuan dalam biomarker prediktif berbasis ML dalam indikasi selain onkologi dengan menggunakan berbagai jenis data masukan. Tasaki et al.[72] menerapkan pendekatan ML pada data multi-omik untuk lebih memahami respons obat untuk pasien dengan rheumatoid arthritis. Pare dkk.[73] mengembangkan kerangka kerja ML baru berdasarkan pohon regresi yang ditingkatkan gradiennya untuk membangun skor risiko poligenik untuk memprediksi sifat-sifat yang kompleks. Diuji pada kumpulan data UK Biobank, model berbasis SNP mereka mampu menjelaskan 46,9% dan 32,7% dari keseluruhan varians poligenik untuk tinggi badan dan BMI. Selain itu, Khera dkk.[74] mengembangkan skor poligenik seluruh genom untuk mengidentifikasi individu yang berisiko tinggi terkena penyakit arteri koroner, fibrilasi atrium, diabetes tipe 2, penyakit radang usus, dan kanker payudara.

Evolusi yang cepat dari teknologi pengurutan RNA sel tunggal telah digunakan untuk pengelompokan gen dan penemuan biomarker spesifik sel. Teknik pengurutan RNA sel tunggal telah digunakan untuk mengidentifikasi jenis sel baru, membedakan status sel, melacak garis keturunan perkembangan dan mengintegrasikan profil ekspresi dengan resolusi spasial sel. Namun, tantangan yang belum terpecahkan adalah pengurangan pengukuran ekspresi gen dari puluhan ribu sel ke ruang berdimensi rendah, biasanya dua atau tiga variabel. Ding dkk.[75] mengembangkan model generatif probabilistik, scvis, untuk mereduksi ruang berdimensi tinggi ke struktur berdimensi rendah dalam data ekspresi gen sel tunggal dengan estimasi ketidakpastian. Alat ini kemudian digunakan untuk menganalisis empat set data sekuensing RNA sel tunggal dan menghasilkan representasi 2D dari data sekuensing RNA sel tunggal multi-dimensi yang dapat ditafsirkan untuk mengidentifikasi jenis sel dengan kuat. Selain itu, Rashid et al.[76] telah menggunakan vari- ational autoencoders (VAE) untuk mengubah data sekuensing RNA sel tunggal menjadi ruang fitur tersandi laten yang secara lebih efisien membedakan antara subpopulasi tumor yang tersembunyi. Analisis ruang fitur yang dikodekan mengungkapkan subpopulasi sel dan hubungan evolusioner di antara mereka. Metode ini sepenuhnya tanpa pengawasan dan membutuhkan pra-pemrosesan data yang minimal. Selain itu, metode ini toleran terhadap penurunan ekspresi gen dalam set data sekuensing RNA sel tunggal. Wang dan [Gu77] mengusulkan autoencoder variasi dalam untuk data sekuensing RNA sel tunggal (VASC), model generatif multi-layer yang dalam, untuk pengurangan dimensi tanpa pengawasan dan visualisasi

dari data ini. Diuji pada 20 set data, VASC lebih unggul dan memiliki kompatibilitas set data yang lebih luas daripada beberapa metode reduksi dimensi yang mutakhir seperti ZIFA78 dan SIMLR79.

Salah satu perkembangan terbaru yang menarik dalam ML adalah peningkatan pesat dalam pemilihan fitur untuk penemuan biomarker. Sebagai contoh, para peneliti menerapkan model DL tanpa pengawasan untuk mengekstrak representasi yang berarti dari modul gen atau klaster sampel80. Way dan Greene81 memperkenalkan model VAE yang dilatih pada data sekuensing RNA pan-kanker The Cancer Genome Atlas (TCGA) dan mengidentifikasi pola khusus dalam fitur yang dikodekan oleh VAE. Beck et al.[82] melakukan analisis gambar dan integrasi data dengan ekspresi gen dan data proteomik untuk meningkatkan identifikasi karsinoma sel skuamosa paru. Nirschl et al.[83] menunjukkan bahwa model CNN dapat memprediksi kemungkinan gagal jantung dari sampel biopsi endomiokard dengan lebih baik (AUC = 0,97) dibandingkan dengan dua ahli patologi jantung yang terlatih (AUC = 0,73 dan 0,75).

Dalam semua contoh ini, agar biomarker prediktif yang dihasilkan ML menjadi lebih sukses, ada tujuh masalah utama yang masih perlu diatasi. Setidaknya beberapa masalah ini menyangkut kemampuan interpretasi pengklasifikasi, yang dianggap oleh setidaknya beberapa pengguna akhir sebagai hal yang penting untuk adopsi klinis. Salah satu masalah utama lainnya adalah kebutuhan untuk memvalidasi pendekatan ini dalam konteks kumpulan data multi-situs, multi-institusi untuk menunjukkan kemampuan generalisasi dari pendekatan tersebut. Komunitas penelitian secara aktif menangani masalah ini dan membuat kemajuan pesat, termasuk penerapan pendekatan objektif dan langkah-langkah untuk pelatihan model dan optimasi parameter84, interpretasi model dan ekstraksi wawasan biologis85, dan reproduktifitas model86.

Patologi komputasi. Patologi adalah bidang deskriptif, karena ahli patologi menafsirkan apa yang terlihat pada slide kaca dengan inspeksi visual. Analisis slide kaca ini memberikan banyak sekali informasi, seperti jenis sel yang ada dalam jaringan dan konteks spasialnya. Interaksi antara tumor dan sel imun dalam lingkungan mikro tumor semakin penting dalam studi imuno-onkologi dan tidak dapat ditangkap oleh teknologi lain.

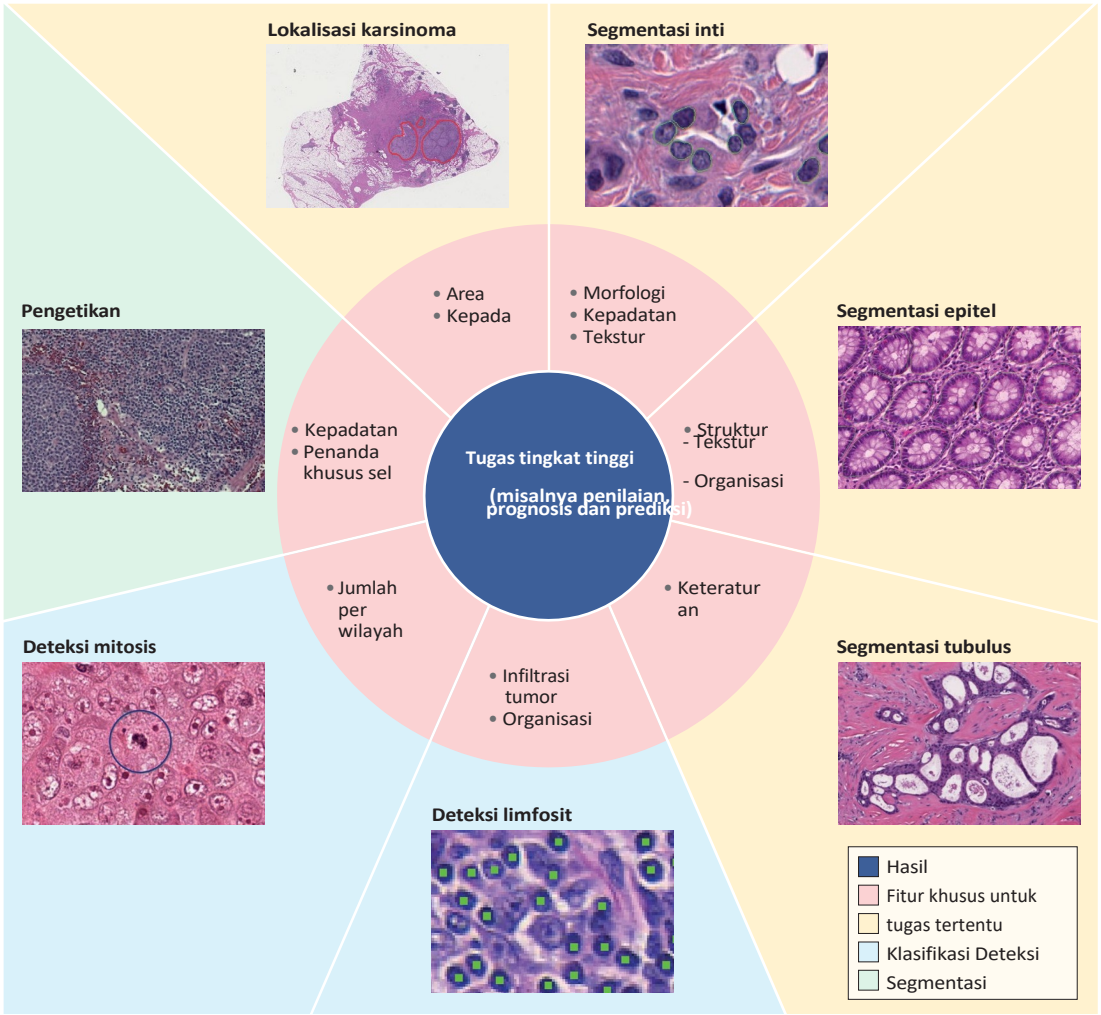
Perusahaan farmasi perlu memahami bagaimana pengobatan obat memengaruhi jaringan dan sel tertentu dan perlu menguji ribuan senyawa sebelum memilih kandidat untuk uji klinis. Selain itu, seiring dengan bertambahnya jumlah uji klinis, penemuan biomarker baru akan semakin penting untuk mengidentifikasi pasien yang akan merespons terapi tertentu. Peningkatan penggunaan patologi komputasi yang memungkinkan penemuan biomarker baru dan menghasilkannya dengan cara yang lebih tepat, dapat direproduksi, dan dengan hasil yang tinggi pada akhirnya akan memangkas waktu pengembangan obat dan memungkinkan pasien mendapatkan akses yang lebih cepat ke terapi yang bermanfaat.

Sebelum DL, algoritme untuk analisis gambar jaringan sering kali terinspirasi secara biologis dalam kolaborasi dengan ahli patologi dan mengharuskan para ilmuwan komputer untuk membuat fitur deskriptif agar komputer dapat mengklasifikasikan jenis jaringan atau sel tertentu. Studi-studi ini ditujukan untuk mengidentifikasi deskriptor morfologi dalam

gambar yang diwarnai dengan haemotoxylin dan eosin (H&E). Morfometri nuklir adalah salah satu implementasi paling awal dari patologi komputasi, yang menunjukkan kemampuan untuk menentukan hubungan antara fitur yang dihasilkan komputer dan prognosis[87]. Beck et al.[88] melihat sel dalam konteks lokasi spasial mereka di dalam stroma tumor yang membulat dan menunjukkan hubungan antara fitur stroma dan kelangsungan hidup pada kanker payudara. Lee et al.[89] juga telah menunjukkan bahwa analisis komputasi jaringan jinak yang berdekatan dengan tumor pada kanker prostat dapat mengungkapkan informasi yang biasanya diabaikan oleh ahli patologi tetapi dikaitkan dengan kelangsungan hidup bebas perkembangan. Baru-baru ini, Lu dkk. menunjukkan bahwa fitur yang menggambarkan bentuk nuklir dan orientasi nuklir sangat terkait dengan kelangsungan hidup pada kanker mulut[90] dan kanker payudara reseptor estrogen positif tahap awal[91]. Dalam banyak kasus, ketersediaan pewarnaan imunohistokimia, yang menggunakan antibodi untuk menargetkan protein spesifik dalam gambar dan menandai jenis sel dan jaringan tertentu, menghindari kebutuhan untuk deteksi sel dan jaringan dengan morfologi dan dengan demikian memungkinkan dihasilkannya data yang canggih tanpa menggunakan alat DL. Namun, dalam kasus imuno-onkologi, ML memungkinkan pembuatan fitur dengan throughput tinggi yang menggambarkan hubungan spasial untuk ribuan sel, tugas yang tidak mungkin dilakukan oleh ahli patologi. Peningkatan dalam deteksi sel dan jaringan individu melalui metode DL memungkinkan pengukuran yang sangat tepat dari lingkungan mikro tumor, sehingga fitur-fitur yang sangat beragam yang menggambarkan hubungan spasial antara sel dan struktur jaringan sekarang dapat diukur dalam skala besar (Gbr. 5).

Dalam sebuah studi oleh Mani et al.[92], beberapa penanda untuk limfosit
fosit digunakan untuk memahami heterogenitas populasi ini pada kanker payudara. Giraldo dkk.[93] meneliti interaksi sel-sel dan menunjukkan bahwa, dengan menggunakan kepadatan sel dan lokasi relatif sel $^{PD1+}$ dan $^{CD8+}$, mereka dapat mengidentifikasi pasien dengan karsinoma sel Merkel yang akan merespons pembrolizumab. Trade-off untuk jenis percobaan ini adalah bahwa mereka menggunakan banyak jaringan, biasanya membutuhkan slide tambahan untuk setiap pewarnaan; namun, ratusan atau ribuan fitur dapat diperiksa, dan jumlah kemungkinan tindakan antar sel meningkat dengan setiap pewarnaan yang digunakan. Dalam kasus seperti itu, kombinasi pemilihan fitur dan metode ML digunakan untuk menentukan kombinasi yang mungkin dapat memprediksi respons terapeutik.

Penerapan CNN pada gambar patologi bekerja dengan baik karena ada sejumlah besar piksel yang layak yang dapat digunakan untuk pelatihan dari satu biopsi atau reseksi. Dengan contoh yang cukup banyak, algoritma DL dapat dirancang untuk mempelajari fitur secara otomatis untuk berbagai macam tugas klasifikasi[94]. Sebagai contoh, jaringan saraf konvolusi multi-skala (M-CNN) digunakan dalam pendekatan pembelajaran yang diawasi untuk fenotipe gambar seluler berkonten tinggi[9] dalam satu langkah dibandingkan dengan beberapa langkah khusus yang independen. Dengan hanya menggunakan nilai intensitas piksel dari gambar untuk mengubah gambar tersebut menjadi fenotipe, pendekatan ini menghasilkan klasifikasi yang lebih akurat secara keseluruhan tentang efek perlakuan senyawa pada berbagai konsentrasi. Banyak tantangan analisis gambar telah berhasil menggunakan metode DL untuk mengidentifikasi area dalam tumor kanker[95-98], tubulus[99],



Gbr. 5 | **tugas-tugas patologi komputasi untuk aplikasi pembelajaran mesin.** Kerangka kerja pembelajaran mendalam dapat menggantikan
fitur buatan tangan tradisional dalam beberapa tugas pengenalan gambar patologi dasar (seperti segmentasi nukleus, epitel atau tubulus, deteksi limfosit, deteksi mitosis atau klasifikasi tumor) dengan menggunakan segmentasi gambar (latar belakang kuning), deteksi fitur spesifik (latar belakang biru) atau deteksi sekumpulan fitur yang digunakan untuk klasifikasi (latar belakang hijau). Pengenalan didasarkan pada fitur-fitur khusus tugas yang ditunjukkan di wilayah merah muda dan dapat menghasilkan prognosis atau prediksi penyakit yang lebih akurat.

aktivitas mitosis100 dan limfosit101,102 pada kanker payudara dan paru-paru.

Selain gambar patologi, DL juga dapat memfasilitasi integrasi modalitas informasi lainnya. DL juga dapat digunakan untuk mempercepat akuisisi data magnetic resonance imaging (MRI)[103] atau mengurangi dosis radiasi yang diperlukan untuk pencitraan computed tomography (CT)[104]. Dengan kualitas pencitraan yang lebih baik termasuk resolusi temporal dan spasial serta rasio sinyal terhadap noise yang tinggi, kinerja analisis gambar juga dapat meningkat dalam aplikasi seperti kuantifikasi gambar, deteksi jaringan abnormal, stratifikasi pasien, dan diagnosis atau prediksi penyakit. Penelitian terbaru lainnya105 menunjukkan kemampuan untuk menggunakan kerangka kerja DL awal untuk memprediksi keberadaan gen tertentu yang bermutasi dari gambar tumor paru-paru yang diwarnai H&E.

Namun demikian, meskipun DL terus unggul dalam banyak tugas analisis gambar tertentu, dalam praktiknya, kombinasi DL dan algoritme analisis gambar tradisional diterapkan pada sebagian besar set masalah. Hal ini dilakukan untuk beberapa alasan

alasan. Pertama, meskipun DL telah menunjukkan kemampuannya untuk menyamai atau mengungguli manusia dalam masalah yang sangat spesifik (seperti mendeteksi glomerulus), namun DL masih belum menjadi alat analisis gambar untuk tujuan umum yang hebat. Waktu pengembangannya masih lama karena kurangnya fleksibilitas. Ada juga kelangkaan label ahli yang tersedia untuk tugas klasifikasi tertentu, karena ini mahal untuk dibuat. Pendekatan untuk mengurangi hal ini termasuk menggunakan pewarnaan imunohisokimia untuk memberikan informasi tambahan kepada ahli patologi untuk sampel di mana anotasi sulit [dilakukan106] serta upaya untuk meningkatkan ketersediaan anotasi ahli yang dikurasi dengan baik untuk kasus-kasus yang digunakan secara luas (sel kanker versus sel normal), yang merupakan tugas komunitas yang sedang berlangsung.

Tantangan lainnya adalah masalah transparansi. Metode DL dikenal dengan pendekatan kotak hitamnya. Alasan yang mendasari di balik keputusan untuk tugas klasifikasi tidak jelas. Untuk pengembangan obat, penting untuk memahami mekanisme, dan memiliki output yang dapat ditafsirkan dapat berguna untuk menemukan tidak hanya potensi baru

target obat tetapi juga biomarker potensial baru untuk menentukan respons terapeutik. Pembuatan lebih banyak fitur buatan tangan diperlukan untuk meningkatkan kepercayaan dalam interpretasi.

Tantangan lebih lanjut adalah ukuran sampel yang besar yang dibutuhkan dalam uji klinis untuk menerapkan DL secara langsung untuk menyimpulkan respons terapeutik. DL biasanya membutuhkan puluhan ribu atau bahkan ratusan ribu contoh untuk dipelajari, dan uji klinis biasanya tidak menghasilkan contoh yang cukup. Dalam kasus tertentu, dimungkinkan untuk menggabungkan data di seluruh uji klinis, tetapi bias mungkin ada yang dapat membuat hasilnya lebih sulit untuk ditafsirkan.

Contoh integrasi yang berhasil dari DL dan alur kerja analisis gambar tradisional termasuk karya Saltz dkk.[101] dan Corredor dkk.[102], di mana CNN digunakan untuk mendeteksi limfosit dalam jaringan yang diwarnai dengan H&E dan fitur berbasis grafik sub-sekuen diekstraksi untuk memprediksi respons penyakit. Hal ini kemungkinan akan menjadi peran umum untuk DL dalam waktu dekat, karena kemampuannya yang unggul dalam mendeteksi sel dan jaringan dapat menggantikan segmentasi tradisional dan algoritme deteksi nuklir, dan fitur-fitur yang dapat diinterpretasikan selanjutnya dapat diterapkan untuk memberikan konteks spasial pada fitur-fitur ini.

Pandangan

Pendekatan ML dan perkembangan terbaru dalam DL memberikan banyak peluang untuk meningkatkan efisiensi di seluruh jalur penemuan dan pengembangan obat. Dengan demikian, kami berharap dapat melihat peningkatan jumlah aplikasi untuk masalah yang terdefinisi dengan baik di seluruh industri dalam beberapa tahun mendatang. Dengan data yang tersedia menjadi 'lebih besar', setidaknya dalam arti lebih menyeluruh mencakup variabilitas yang relevan dari seluruh ruang data, dan ketika komputer menjadi semakin kuat, algoritma ML akan secara sistematis menghasilkan output yang lebih baik, dan aplikasi baru yang menarik diharapkan akan mengikuti. Hal ini telah dicontohkan dengan jelas pada bagian sebelumnya, di mana kami telah menjelaskan beberapa aplikasi ML untuk identifikasi dan validasi target, desain dan pengembangan obat, identifikasi biomarker dan patologi untuk diagnosis penyakit dan prognosis terapi di klinik. Metode-metode ini juga sedang diterapkan dalam lingkungan perawatan kesehatan, yang, jika digabungkan dengan penemuan obat, dapat menghasilkan kemajuan yang signifikan dalam pengobatan yang dipersonalisasi[107]. ML juga telah diterapkan pada catatan kesehatan elektronik[108] dan bukti dunia nyata untuk meningkatkan hasil uji klinis dan mengoptimalkan proses penilaian kelayakan uji klinis. Sebagai contoh, sebuah penelitian baru-baru ini menunjukkan bahwa DNN adalah pendekatan yang sangat kompetitif untuk secara otomatis mengekstraksi informasi yang berguna dari rekam medis elektronik untuk diagnosa dan klasifikasi penyakit[109]. Beberapa penelitian menunjukkan bahwa model ML dalam catatan kesehatan elektronik dapat mengungguli model konvensional dalam memprediksi prognosis[110]. ML juga dapat diterapkan pada data yang sekarang berasal dari sensor dan perangkat yang dapat dikenakan untuk memahami penyakit dan mengembangkan perawatan, terutama dalam ilmu saraf[111]. Gkotsis dkk.[112] menerapkan pendekatan DL untuk mengkarakterisasi kondisi kesehatan mental pada data media sosial yang tidak terstruktur, yang merupakan data yang sulit tugas untuk pendekatan TPPU tradisional.

Seperti yang ditunjukkan pada Gbr. 1, pendekatan ML mulai umum digunakan dalam berbagai langkah penemuan

dan jalur pengembangan oleh perusahaan farmasi. Implementasi metode ML yang meluas ini memiliki beberapa masalah yang diketahui penting. Masalah yang umum terjadi pada jaringan syaraf tiruan yang terlatih adalah kurangnya inter-pretabilitas, yaitu kesulitan untuk mendapatkan penjelasan yang sesuai dari jaringan syaraf tiruan yang terlatih mengenai bagaimana jaringan syaraf tiruan tersebut sampai pada hasilnya. Jika sistem digunakan untuk mendiagnosis penyakit seperti melanoma, misalnya, berdasarkan gambar medis, kurangnya kemampuan interpretasi ini dapat menghalangi para ilmuwan, badan pengatur, dokter, dan pasien, bahkan dalam situasi di mana jaringan saraf bekerja lebih baik daripada pakar manusia. Akankah pasien lebih mempercayai diagnosis ML daripada diagnosis ahli manusia? Meskipun tidak terlalu dramatis, situasi yang sama dapat terjadi dalam desain obat. Akankah perusahaan farmasi mempercayai jaringan syaraf tiruan untuk memilih molekul kecil untuk dimasukkan ke dalam portofolio dan investasi mereka untuk dikembangkan ke klinik, tanpa penjelasan yang jelas mengapa jaringan syaraf tiruan memilih molekul ini? Selain itu, mungkin ada masalah aplikasi paten dengan penemu jika senyawa telah dirancang oleh algoritma komputer. Bagaimanapun, hasil ML harus dianggap hanya sebagai hipotesis atau titik awal yang menarik yang kemudian dikembangkan lebih lanjut dalam studi oleh para peneliti. Eksperimen pelengkap yang memvalidasi hasil ML akan membantu membangun kepercayaan pada pendekatan dan output, tetapi badan pengatur belum mengklarifikasi pandangan mereka tentang kurangnya interpretabilitas untuk penggunaan ML secara klinis. Namun, bahkan di luar masalah kepercayaan, kurangnya interprestasi dari pendekatan-pendekatan tersebut membuat pendekatan-pendekatan ini lebih sulit untuk dipecahkan ketika secara tak terduga gagal pada set data baru yang belum pernah ada sebelumnya.

Masalah penting lainnya untuk jaringan syaraf adalah pengulangan, yang muncul karena keluaran ML sangat bergantung pada nilai awal atau bobot parameter jaringan atau bahkan urutan contoh pelatihan yang diberikan kepada jaringan, karena semuanya biasanya dipilih secara acak. Apakah jaringan akan selalu memilih target penyakit yang sama dengan menggunakan data ekspresi yang sama sebagai input? Apakah struktur obat yang diusulkan oleh metode ML akan selalu sama? Kurangnya pengulangan ini sangat bermasalah untuk identifikasi biomarker, seperti yang terlihat dalam situasi di mana alat yang berbeda menghasilkan biomarker prognosis yang berbeda untuk kanker payudara berdasarkan tanda tangan ekspresi molekuler[113]. Fakta bahwa metode ML yang berbeda dapat memberikan hasil yang berbeda akan menambah ketidakpastian pada penerapan metode ini dalam skala besar. Beberapa solusi untuk masalah interpretabilitas dan pengulangan telah diusulkan. Solusi tersebut biasanya berpusat pada penggunaan algoritma yang lebih kompleks atau lebih memakan waktu atau rata-rata hasil dari beberapa model jaringan, tetapi hal ini dapat dilihat sebagai penambahan satu hasil lagi pada berbagai hasil yang sudah ada. Hal penting lainnya yang perlu dipertimbangkan adalah ketersediaan data yang berkualitas tinggi, akurat, dan terkurasi dalam jumlah yang besar untuk melatih dan mengembangkan model ML. Persyaratan untuk jumlah dan akurasi yang diinginkan tergantung pada kompleksitas tipe data dan pertanyaan yang harus diselesaikan. Oleh karena itu, biaya yang dibutuhkan untuk menghasilkan kumpulan data ini bisa jadi mahal. Konsorsium pra-kompetitif perusahaan farmasi dan institusi akademik yang menggunakan standar data yang sesuai dan memiliki

Kerangka kerja operasional dan data terbuka yang diperlukan dapat menjadi bagian dari solusi untuk memenuhi kebutuhan data ini. Banyak jenis data yang digunakan selama penemuan obat masih jauh dari komprehensif. Sebagai contoh, pengetahuan tentang semua lipatan dan struktur protein tidak lengkap, dan cakupan ruang data juga tidak lengkap. Dengan demikian, aplikasi di mana struktur-struktur ini diprediksi, bahkan jika banyak kemajuan telah dibuat, belum sebaik di bidang lain. Hal yang sama berlaku untuk prediksi reaksi yang terlibat dalam sintesis molekul kecil yang seluruh ruang kimianya tidak diketahui.

Kurasi data adalah kunci untuk penyediaan data yang dapat digunakan kembali dan dapat dipercaya, dan bisa jadi mahal dalam hal waktu dan keterampilan yang dibutuhkan. Kurasi biologi - ekstraksi informasi biologi dari literatur ilmiah dan integrasinya ke dalam basis data - berada di antara seni dan ilmu pengetahuan114, yang membutuhkan kombinasi keterampilan komputasi dengan keahlian biologi dan domain yang mendalam. Upaya kolaboratif untuk mengembangkan sumber daya data bersama dan metadata (label) dapat menjadi cara agar data berkualitas tinggi dalam domain publik dapat lebih tersedia. Hal ini juga termasuk metadata dari program penemuan obat yang berhasil maupun yang gagal yang dapat memungkinkan pendekatan prediksi dan penentuan faktor-faktor yang dapat mengurangi gesekan dalam pengembangan obat. Kolaborasi pra-kompetitif yang lebih banyak juga diperlukan untuk mengumpulkan dan menghasilkan sumber daya data yang besar dari kumpulan data bioaktif perusahaan dari senyawa-senyawa yang sedang diselidiki serta data uji klinis historis.

Keterbatasan lain dalam penerapan model ML adalah penggunaannya untuk memprediksi paradigma alternatif. Karena seluruh premis ML bergantung pada penggunaan data pelatihan untuk menghasilkan model yang sesuai, model ML hanya dapat memprediksi dalam kerangka kerja pelatihan yang diketahui. data. Dalam kimia medisinal, misalnya, desain senyawa dengan mekanisme kerja alternatif, seperti makrosiklus, penghambat interaksi protein-protein, atau PROTAC, mungkin hanya dapat dilakukan dengan metode tradisional.

Selain data dan model, pelatihan para peneliti yang memahami ilmu farmasi serta ilmu komputer, statistik komputasi dan ML statistik dan mahir dalam menggunakan metode-metode ini perlu dipercepat. Kompetisi seperti DREAM Challenges (lihat tautan terkait), yang telah menunjukkan bahwa komposisi tim merupakan faktor dalam kinerja, juga dapat berguna untuk menarik bakat dan memajukan pengembangan metodologi. Namun, aplikasi harus berhasil dalam pengaturan klinis untuk memotivasi investasi lebih lanjut dari perusahaan farmasi dan teknologi besar.

Algoritme ML, termasuk metode DL, telah memungkinkan pemanfaatan AI dalam lingkungan industri dan kehidupan sehari-hari. Dampak dari metode ML di semua bidang penemuan obat dan perawatan kesehatan sudah dirasakan, terutama dalam analisis data omics dan pencitraan. Algoritma ML juga berhasil dalam pengenalan suara, NLP, visi komputer, dan aplikasi lainnya. Sebagai contoh, asisten pintar yang diaktifkan di Internet sekarang sudah umum digunakan dan dapat mengirimkan informasi terkait kesehatan dalam bentuk ucapan dan gambar atau video. Pendekatan ML yang diterapkan pada data yang dikumpulkan dari penggabungan teknologi yang diaktifkan di Internet, ditambah dengan data biologis, memiliki potensi untuk secara dramatis meningkatkan kekuatan prediksi algoritme tersebut dan membantu pengambilan keputusan medis tentang manfaat terapeutik, biomarker klinis, dan efek samping terapi.

Diterbitkan secara online

Published online: 11 April 2019

1. Mamoshina, P. dkk. Pembelajaran mesin pada data transkriptomik otot manusia untuk penemuan biomarker dan identifikasi target obat spesifik jaringan. *Depan. Genet.* **9**, 242 (2018).
2. LeCun, Y., Bengio, Y. & Hinton, G. Pembelajaran mendalam. *Nature* **521**, 436 (2015).
3. Chen, H., Engkvist, O., Wang, Y., Olivecrona, M. & Blaschke, T. Munculnya pembelajaran mendalam dalam penemuan obat. *Penemuan Obat. Hari ini* **23**, 1241-1250 (2018).
Artikel ini adalah upaya pertama untuk menyoroti aplikasi terbaru DL dalam penelitian penemuan obat dan merupakan pengantar untuk beberapa arsitektur DL yang **populer.**
4. Hinton, G. Pembelajaran mendalam - teknologi yang berpotensi mengubah perawatan kesehatan. *JAMA* **320**, 1101-1102 (2018).
5. Wong, CH, Siah, KW & Lo, AW Estimasi tingkat keberhasilan uji klinis dan parameter terkait. Biostatistik https://doi.org/10.1093/biostatistics/kxx069 (2018).
6. Jeon, J. dkk. Pendekatan sistematis untuk mengidentifikasi target obat kanker baru menggunakan pembelajaran mesin, desain inhibitor, dan skrining dengan hasil tinggi. *Genome Med.* **6**, 57 (2014).
7. Ferrero, E., Dunham, I. & Sanseau, P. Prediksi in silico target terapi baru menggunakan data asosiasi gen-penyakit . *J. Transl Med.* **15**, 182 (2017).
8. Riniker, S., Wang, Y., Jenkins, J. & Landrum, G. Menggunakan informasi dari layar historis berkinerja tinggi untuk memprediksi senyawa aktif. *J. Chem. Inf. Model.* **54**, 1880-1891 (2014).
9. Godinez, WJ, Hossain, I., Lazic, SE, Davies, JW & Zhang, X. Jaringan saraf konvolusi multi-skala untuk fenotipe gambar seluler konten tinggi. *Bioinformatika* **33**, 2010-2019 (2017).
10. Olsen, T. dkk. Kinerja diagnostik algoritma pembelajaran mendalam yang diterapkan pada tiga diagnosis dalam dermatopatologi. *J. Pathol. Inform.* **9**, 32-32 (2018).
11. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Putus sekolah: cara sederhana untuk mencegah jaringan syaraf agar tidak overfitting. *J. Mach. Learn. Res.* **15**, 1929-1958 (2014).
12. Jiao, Y. & Pufeng, D. Ukuran kinerja dalam mengevaluasi prediktor bioinformatika berbasis pembelajaran mesin untuk klasifikasi. *Quant. Biol.* **4**, 320 (2016).
13. Czodrowski, P. Mengandalkan kappa. *J. Comput. Dibantu Mol. Des.* **28**, 1049-1055 (2014).
14. Rifaioglu, AS dkk. Aplikasi terbaru dari pembelajaran mendalam dan kecerdasan mesin pada penemuan obat in silico: metode, alat, dan basis data. *Ringkasan. Bioinform.* https://doi.org/10.1093/bib/bby061 (2018).
15. Hinton, G. E. & Salakhutdinov, R. R. Mengurangi dimensi data dengan jaringan syaraf. *Science* **313**, 504 (2006).
16. Koscielny, G. dkk. Target terbuka: platform untuk identifikasi dan validasi target terapeutik. *Asam Nukleat Res.* **45**, D985-D994 (2017).
17. Costa, PR, Acencio, ML & Lemke, N. Pendekatan pembelajaran mesin untuk prediksi genom di seluruh genom gen manusia yang tidak sehat dan yang dapat dibius berdasarkan data tingkat sistem . *BMC Genomics* **11**, S9-S9 (2010).
18. Ament, SA dkk. Jaringan regulasi transkripsional yang mendasari perubahan ekspresi gen pada penyakit Huntington. *Mol. Sistem Biol.* **14**, e7435 (2018).
19. Bravo, A., Pinero, J., Queralt-Rosinach, N., Rautschka,

M. & Furlong, LI Ekstraksi hubungan antara gen dan penyakit dari teks dan analisis data berskala besar: implikasi untuk penelitian translasi. *BMC Bioinformatika* **16**, 55 (2015).
20. Kim, J., Kim, J.-j. & Lee, H. Analisis hubungan gen penyakit dari abstrak Medline oleh DigSee. *Sci. Rep.* **7**, 40154 (2017).
21. Leung, MK, Xiong, HY, Lee, LJ & Frey, BJ Pembelajaran mendalam tentang kode penyambungan yang diatur oleh jaringan. *Bioinformatika* **30**, i121-i129 (2014).
22. Jha, A., Gazzara, MR & Barash, Y. Model mendalam integratif untuk penyambungan alternatif. *Bioinformatika* **33**, i274-i282 (2017).
23. Vaquero-Garcia, J. dkk. Pandangan baru tentang kompleksitas dan regulasi transkriptom melalui lensa variasi penyambungan lokal. *eLife* **5**, e11752 (2016).
24. Sotillo, E. dkk. Konvergensi mutasi yang didapat dan penyambungan alternatif CD19 memungkinkan resistensi terhadap imunoterapi CART-19. *Cancer Discov.* **5**, 1282-1295 (2015).
25. Rohacek, AM dkk. Mutasi ESRP1 menyebabkan gangguan pendengaran karena cacat pada penyambungan alternatif yang mengganggu perkembangan koklea. *Dev. Sel* **43**, 318-331 (2017).
26. Xiong, HY dkk. Penyambungan RNA. Kode penyambungan manusia mengungkap wawasan baru tentang faktor penentu genetik penyakit. *Science* **347**, 1254806 (2015).
Artikel ini menjelaskan model komputasi berdasarkan DL yang memprediksi regulasi splicing untuk setiap urutan mRNA dan telah diterapkan pada lebih dari setengah juta varian urutan splicing mRNA manusia.
Ribuan mutasi penyebab penyakit yang diketahui telah diidentifikasi serta gen baru yang terkait dengan **penyakit.**
27. Iorio, F. dkk. Lanskap interaksi farmakogenomik pada kanker. *Cell* **166**, 740-754 (2016). **Makalah ini menerapkan ML pada data dari mutasi somatik, perubahan jumlah salinan, metilasi DNA dan ekspresi gen dari 1.000 garis sel kanker untuk memodelkan respons obat dari garis sel dan menunjukkan pentingnya fitur genomik untuk prediksi.**
28. Tsherniak, A. dkk. Mendefinisikan peta ketergantungan kanker . *Sel* **170**, 564-576 (2017).
29. McMillan, EA dkk. Pendekatan kimia pertama untuk nominasi pengobatan yang dipersonalisasi pada kanker paru-paru. *Cell* **173**, 864-878 (2018).

30. Al-Lazikani, B. et al. dalam *Bioinformatika - Dari Genomes to Therapies* Ch. 36 (Wiley-VCH, 2008).
31. Nayal, M. & Honig, B. Tentang sifat rongga pada permukaan protein: aplikasi untuk identifikasi tempat pengikatan obat. *Protein* **63**, 892-906 (2006). **Artikel ini menjelaskan pengklasifikasi untuk mengidentifikasi rongga pengikatan obat berdasarkan atribut fisikokimia, struktural dan geometris protein**.
32. Li, Q. & Lai, L. Prediksi target obat potensial berdasarkan sifat sekuen sederhana. *BMC Bioinformatika* **8**, 353 (2007).
33. Bakheet, TM & Doig, AJ Sifat dan identifikasi target obat protein manusia. *Bioinformatika* **25**, 451-457 (2009).
34. Wang, Q., Feng, Y., Huang, J., Wang, T. & Cheng, G. Kerangka kerja baru untuk identifikasi protein target obat: menggabungkan penyandi otomatis bertumpuk dengan mesin vektor pendukung yang bias. *PLOS ONE* **12**, e0176486 (2017).
35. Kandoi, G., Acencio, ML & Lemke, N. Prediksi protein yang dapat diobati menggunakan pembelajaran mesin dan biologi sistem: tinjauan singkat. *Front. Fisiol.* **6**, 366-366 (2015).
36. Nelson, MR dkk. Dukungan bukti genetik manusia untuk indikasi obat yang disetujui. *Nat. Genet.* **47**, 856-860 (2015).
37. Morgan, P. dkk. Dampak kerangka kerja lima dimensi terhadap produktivitas R&D di AstraZeneca. *Nat. Rev. penemuan obat.* **17**, 167-181 (2018).
38. Rouillard, A. D., Hurle, M. R. & Agarwal, P. Interogasi sistematis terhadap data Omic yang beragam mengungkapkan fitur transkriptomik yang dapat ditafsirkan, kuat, dan dapat digeneralisasi dari target terapi yang berhasil secara klinis. *PLOS Comput. Biol.* **14**, e1006142 (2018).
39. Kumar, V., Sanseau, P., Simola, DF, Hurle, MR & Agarwal, P. Analisis sistematis target obat mengkonfirmasi ekspresi dalam jaringan yang relevan dengan penyakit. *Sci. Rep.* **6**, 36205 (2016).
40. Ramsundar, B. dkk. Apakah pembelajaran mendalam multitask praktis untuk farmasi? *J. Chem. Inf. Model.* **57**, 2068-2076 (2017).
41. Ma, J., Sheridan, RP, Liaw, A., Dahl, GE & Svetnik, V. Jaring saraf dalam sebagai metode untuk hubungan struktur-aktivitas kuantitatif. *J. Chem. Inf. Model.* **55**, 263-274 (2015).
42. Barati Farimani, A., Feinberg, E. & Pande, V. Jalur pengikatan opiat ke reseptor μ-opioid yang diungkapkan oleh pembelajaran mesin . *Biofisika. J.* **114**, 62a - 63a (2018).
43. Wu, Z. dkk. MoleculeNet: tolok ukur untuk pembelajaran mesin molekuler. *Chem. Sci.* **9**, 513-530 (2018).
44. Segler, MHS, Preuss, M. & Waller, MP Merencanakan sintesis kimia dengan jaringan saraf tiruan dan AI simbolik. *Nature* **555**, 604 (2018). **Makalah ini menjelaskan pendekatan yang sangat menyeluruh untuk analisis retrosintesis. Para penulis menunjukkan bahwa metode mereka dapat bersaing dengan retrosintesis yang dilakukan oleh ahli kimia berpengalaman yang ahli dalam bidang ini**.
45. Olivecrona, M., Blaschke, T., Engkvist, O. & Chen, H. Desain de-novo molekuler melalui penguatan mendalam pembelajaran. *J. Cheminform.* **9**, 48 (2017).
46. Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A. & Zhavoronkov, A. druGAN: generatif tingkat lanjut model autoencoder adversarial untuk generasi de novo molekul baru dengan sifat molekul yang diinginkan di silico. *Mol. Pharm.* **14**, 3098-3104 (2017).
47. Smith, JS, Roitberg, AE & Isayev, O. Mengubah penemuan obat komputasi dengan pembelajaran mesin dan AI. *ACS Med. Chem. Lett.* **9**, 1065-1069 (2018).
48. Lenselink, EB dkk. Di luar hype: jaringan saraf dalam mengungguli metode yang sudah mapan menggunakan set tolok ukur bioaktivitas ChEMBL. *J. Cheminform.* **9**, 45 (2017).
49. Gaulton, A. dkk. Basis data ChEMBL pada tahun 2017. *Asam Nukleat Res.* **45**, D945-D954 (2017).
50. Ramsundar, B. dkk. Jaringan multitask secara masif untuk penemuan obat. Tersedia di *arXiv* https://arxiv.org/ abs/1502.02072 (2015).
51. Gutlein, M. & Kramer, S. Sidik jari melingkar yang difilter meningkatkan kinerja prediksi atau runtime dengan tetap mempertahankan kemampuan interpretasi. *J. Cheminform.* **8**, 60 (2016).
52. Mayr, A. dkk. Perbandingan skala besar metode pembelajaran mesin untuk prediksi target obat pada ChEMBL. *Chem. Sci.* **9**, 5441-5451 (2018). **Makalah penelitian ini menjelaskan metodologi yang digunakan oleh para pemenang dari hampir semua kategori dari Tox21 Challenge.**
53. Keiser, MJ dkk. Menghubungkan farmakologi protein dengan kimia ligan . *Nat. Biotechnol.* **25**, 197 (2007).
54. Preuer, K., Renz, P., Unterthiner, T., Hochreiter, S. & Klambauer, G. Fréchet ChemNet Distance: sebuah metrik untuk model generatif untuk molekul dalam penemuan obat. *J. Chem. Inf. Model.* **58**, 1736-1741 (2018).
55. Unterthiner, T., Mayr, A., Klambauer, G. & Hochreiter, S. Prediksi toksisitas menggunakan pembelajaran mendalam. Tersedia di *arXiv* https://arxiv.org/abs/1503.01445 (2015).
56. Li, B. dkk. Pengembangan kerangka kerja pemodelan respons obat untuk mengidentifikasi biomarker translasi turunan dari garis sel yang dapat memprediksi hasil pengobatan terhadap erlotinib atau sorafenib. *PLOS ONE* **10**, e0130700 (2015). **Dalam makalah ini, biomarker** prediktif **translasi digunakan untuk menunjukkan bahwa model prediktif dapat dihasilkan dari set data pelatihan praklinis dan kemudian diterapkan pada sampel pasien klinis untuk membuat stratifikasi pasien, menyimpulkan mekanisme kerja obat, dan memilih indikasi penyakit yang sesuai.**
57. van Gool, AJ dkk. Menjembatani kesenjangan inovasi translasi melalui praktik biomarker yang baik. *Nat. Rev. penemuan obat.* **16**, 587-588 (2017).
58. Kraus, VB Biomarker sebagai alat pengembangan obat: penemuan, validasi, kualifikasi, dan penggunaan. *Nat. Rev. Rheumatol.* **14**, 354-362 (2018).
59. Shi, L. dkk. MicroArray Quality Control (MAQC)-II studi tentang praktik umum untuk pengembangan dan validasi model prediktif berbasis microarray. *Nat. Bioteknologi.* **28**, 827-838 (2010).
60. Zhan, F. dkk. Klasifikasi molekuler dari beberapa mieloma. *Blood* **108**, 2020-2028 (2006).
61. Shaughnessy, JD Jr dkk. Model ekspresi gen yang divalidasi dari multiple myeloma berisiko tinggi didefinisikan oleh ekspresi gen yang mengalami deregulasi yang memetakan ke kromosom 1. *Blood* **109**, 2276-2284 (2007).
62. Zhan, F., Barlogie, B., Mulligan, G., Shaughnessy, JD Jr & Bryant, B. Mieloma risiko tinggi: model stratifikasi risiko berbasis ekspresi gen untuk mieloma multipel yang baru didiagnosis yang diobati dengan terapi dosis tinggi adalah prediktif terhadap hasil pada penyakit yang kambuh yang diobati dengan bortezomib agen tunggal atau deksametason dosis tinggi. *Blood* **111**, 968-969 (2008).
63. Decaux, O. dkk. Prediksi kelangsungan hidup pada mieloma multipel berdasarkan profil ekspresi gen mengungkapkan tanda siklus sel dan ketidakstabilan kromosom pada pasien berisiko tinggi dan tanda hiperdiploid pada pasien berisiko rendah: sebuah studi dari Intergroupe Francophone du Myelome. *J. Clin. Oncol.* **26**, 4798-4805 (2008).
64. Mulligan, G. dkk. Profil ekspresi gen dan korelasi dengan hasil dalam uji klinis inhibitor proteasome bortezomib. *Blood* **109**, 3177-3188 (2007).
65. Costello, JC dkk. Upaya komunitas untuk menilai dan meningkatkan algoritme prediksi sensitivitas obat. *Nat. Bioteknologi.* **32**, 1202-1212 (2014). **Makalah ini merupakan upaya untuk mengumpulkan dan mengevaluasi secara objektif berbagai pendekatan ML yang dilakukan oleh tim di seluruh dunia terhadap kumpulan data multi-omik dan berbagai senyawa. Kumpulan data dan hasilnya terus digunakan sebagai tolok ukur untuk** pengembangan **dan validasi metode baru .**
66. Rahman, R., Otridge, J. & Pal, R. IntegratedMRF: kerangka kerja berbasis hutan acak untuk mengintegrasikan prediksi dari berbagai jenis data. *Bioinformatika* **33**, 1407-1410 (2017).
67. Bunte, K., Leppäaho, E., Saarinen, I. & Kaski, S. Analisis faktor kelompok jarang untuk biclustering beberapa sumber data. *Bioinformatika* **32**, 2457-2463 (2016).
68. Huang, C., Mezencev, R., McDonald, JF & Vannberg, F. Algoritma pembelajaran mesin sumber terbuka untuk prediksi terapi obat kanker yang optimal. *PLOS ONE* **12**, e0186906 (2017).
69. Hejase, HA & Chan, C. Meningkatkan prediksi sensitivitas obat dengan menggunakan berbagai jenis data. *CPT Pharmacometrics Syst. Pharmacol.* **4**, e2 (2015).
70. Kim, ES dkk. Uji coba BATTLE: mempersonalisasi terapi untuk kanker paru-paru. *Cancer Discov.* **1**, 44-53 (2011).
71. Boyiadzis, MM dkk. Signifikansi dan implikasi dari persetujuan FDA terhadap pembrolizumab untuk penyakit yang ditentukan oleh biomarker. *J. Imunother. Kanker* **6**, 35 (2018).
72. Tasaki, S. dkk. Pemantauan multi-omik respon obat pada rheumatoid arthritis dalam mengejar remisi molekuler. *Nat. Commun.* **9**, 2755 (2018). Penelitian **ini mengidentifikasi tanda tangan molekuler yang resisten terhadap pengobatan obat dan mengilustrasikan pendekatan multi omics untuk memahami respons obat.**
73. Paré, G., Mao, S. & Deng, W. Q. Sebuah heuristik pembelajaran mesin untuk meningkatkan prediksi skor

gen dari sifat-sifat poligenik . *Sci. Rep.* **7**, 12665 (2017).
74. Khera, AV dkk. Skor poligenik di seluruh genom untuk penyakit umum mengidentifikasi individu dengan risiko setara dengan mutasi monogenik. *Nat. Genet.* **50**, 1219-1224 (2018).
75. Ding, J., Condon, A. & Shah, SP Pengurangan dimensi yang dapat ditafsirkan dari data transkriptom sel tunggal dengan model generatif yang mendalam. *Nat. Commun.* **9**, 2002 (2018).
76. Rashid, S., Shah, S., Bar-Joseph, Z. & Pandya, R. Proyek Dhaka: autoencoder variasional untuk membuka kedok heterogenitas tumor dari data genom sel tunggal. Tersedia di *bioRxiv* https://www.biorxiv.org/content/10.1101/183863v4 (2018).
77. Wang, D. & Gu, J. VASC: reduksi dimensi dan visualisasi data RNA-seq sel tunggal dengan autoencoder variasi dalam. *Genomik Proteomik Bioinformatika* **16**, 320-331 (2017).
78. Pierson, E. & Yau, C. ZIFA: reduksi dimensi untuk analisis ekspresi gen sel tunggal yang tidak meningkat. *Genome Biol.* **16**, 241 (2015).
79. Wang, B., Zhu, J., Pierson, E., Ramazzotti, D. & Batzoglou, S. Visualisasi dan analisis data RNA-seq sel tunggal dengan pembelajaran kemiripan berbasis kernel. *Nat. Metode* **14**, 414 (2017).
80. Tan, J., Hammond, JH, Hogan, DA & Greene, CAO. Integrasi berbasis ADAGE dari data ekspresi gen Pseudomonas aeruginosa yang tersedia untuk umum dengan autoencoder denoising menerangi interaksi mikroba-tuan rumah . *mSystems* **1**, e00025-15 (2016).
81. Way, GP & Greene, CS Mengekstraksi ruang laten yang relevan secara biologis dari transkriptom kanker dengan autoencoder yang bervariasi. *Pac. Symp. Biocomput.* **23**, 80-91 (2018).
82. Casanova, R. dkk. Karakterisasi morfoproteomik fragmentasi karsinoma sel skuamosa paru, penanda histologis peningkatan invasivitas tumor. *Cancer Res.* **77**, 2585-2593 (2017).
83. Nirschl, JJ dkk. Pengklasifikasi deep-learning mengidentifikasi pasien dengan gagal jantung klinis menggunakan gambar seluruh slide jaringan H&E. *PLOS ONE* **13**, e0192726 (2018).
84. Angermueller, C., Pärnamaa, T., Parts, L. & Stegle, O. Pembelajaran mendalam untuk biologi komputasi. *Mol. Syst. Biol.* **12**, 878 (2016).
85. Finnegan, A. & Song, J. S. Metode entropi maksimum untuk mengekstraksi fitur yang dipelajari dari jaringan saraf dalam. *PLOS Comput. Biol.* **13**, e1005836 (2017).
86. Hutson, M. Kecerdasan buatan menghadapi krisis reproduktifitas . *Sains* **359**, 725-726 (2018).
87. Veltri, RW, Partin, AW & Miller, MC Tingkat nuklir kuantitatif (QNG): biomarker berbasis analisis gambar baru dari perubahan struktur nuklir yang relevan secara klinis. *J. Sel. Biokimia. Suppl.* **35**, S151-S157 (2000).
88. Beck, AH dkk. Analisis sistematis morfologi kanker payudara mengungkap fitur stroma yang terkait dengan kelangsungan hidup . *Sci. transl Med.* **3**, 108ra113 (2011).
89. Lee, G. dkk. Bentuk dan arsitektur nuklir di bidang jinak memprediksi kekambuhan biokimiawi pada pasien kanker prostat setelah prostatektomi radikal: temuan awal. *Eur. Urol. Fokus* **3**, 457-466 (2017).
90. Lu, C. dkk. Pengklasifikasi citra berbasis histomorfometri kuantitatif karsinoma sel skuamosa rongga mulut dari morfologi nuklir dapat berisiko membuat stratifikasi pasien untuk kelangsungan hidup spesifik penyakit. *Mod. Pathol.* **30**, 1655-1665 (2017).
91. Lu, C. dkk. Fitur bentuk dan orientasi nuklir dari gambar H&E memprediksi kelangsungan hidup pada kanker payudara reseptor estrogen-positif stadium awal. *Lab. Invest.* **98**, 1438-1448 (2018).
92. Mani, N. L. et al. Penilaian kuantitatif heterogenitas spasial limfosit yang menginfiltrasi tumor pada kanker payudara. *Breast Cancer Res.* **18**, 78 (2016).
93. Giraldo, NA dkk. Hubungan diferensial sel PD-1, PD-L1, dan CD8 + dengan respons terhadap pembrolizumab dan keberadaan virus poliomavirus sel Merkel (MCPyV) pada pasien dengan karsinoma sel Merkel (MCC). *Cancer Res.* **77**, 662 (2017).
94. Janowczyk, A. & Madabhushi, A. Pembelajaran mendalam untuk analisis citra patologi digital: tutorial komprehensif dengan kasus penggunaan terpilih. *J. Pathol. Informat.* **7**, 29 (2016).
Artikel ini adalah ulasan komprehensif pertama tentang DL dalam konteks gambar patologi digital. Makalah ini juga secara sistematis menjelaskan dan menyajikan pendekatan untuk melatih dan memvalidasi pengklasifikasi DL untuk sejumlah masalah berbasis gambar dalam patologi digital, termasuk deteksi sel, segmentasi dan klasifikasi jaringan.
95. Sharma, H., Zerbe, N., Klempert, I., Hellwich, O. & Hufnagl, P. Jaringan saraf convolutional dalam untuk klasifikasi otomatis karsinoma lambung menggunakan gambar slide utuh dalam histopatologi digital. *Comput. Med. Grafik Pencitraan.* **61**, 2-13 (2017).

96. Korbar, B. dkk. Pembelajaran mendalam untuk klasifikasi polip kolorektal pada gambar slide utuh. *J. Pathol. Informat.* **8**, 30 (2017).
97. Bychkov, D. dkk. Analisis jaringan berbasis pembelajaran mendalam memprediksi hasil pada kanker kolorektal. *Sci. Rep.* **8**, 3395 (2018).
98. Cruz-Roa, A. dkk. Deteksi kanker payudara invasif yang akurat dan dapat direproduksi pada gambar slide utuh: Pendekatan Deep Learning untuk mengukur luas tumor.
Sci. Rep. **7**, 46450 (2017).
Ini adalah salah satu makalah pertama yang menerapkan DL untuk mengidentifikasi wilayah kanker payudara pada gambar patologi digital dan menunjukkan bahwa pendekatan algoritmik mengungguli para ahli patologi kanker payudara. Ini adalah salah satu studi pertama yang memiliki kumpulan data kasus yang besar (>600) dengan pelatihan independen dan set validasi .
99. Romo-Bucheli, D., Janowczyk, A., Gilmore, H., Romero, E. & Madabhushi, A. Kuantifikasi inti tubulus otomatis dan korelasi dengan kategori risiko onkotipe DX pada gambar slide utuh kanker payudara ER +. *Sci. Rep.* **6**, 32706 (2016).
Artikel ini menerapkan DL untuk mengidentifikasi keberadaan dan lokasi tubulus dalam gambar patologi payudara
dan kemudian menunjukkan bahwa jumlah tubulus yang terdeteksi berkorelasi dengan penilaian risiko kanker payudara melalui tes genom.
Ini adalah salah satu makalah pertama yang menunjukkan bagaimana DL dapat digunakan untuk membangun asosiasi genotipe-fenotipe.
100. Romo-Bucheli, D., Janowczyk, A., Gilmore, H., Romero, E. & Madabhushi, A. Strategi berbasis pembelajaran mendalam untuk mengidentifikasi dan mengaitkan aktivitas mitosis dengan ekspresi gen yang diturunkan dari kategori risiko pada kanker payudara reseptor estrogen positif. *Sitometri* **A 91**, 566-573 (2017).
101. Saltz, J. dkk. Organisasi spasial dan korelasi molekuler limfosit yang menginfiltrasi tumor menggunakan pembelajaran mendalam pada gambar patologi. *Cell Rep.* **23**, 181-193 (2018).
Penelitian berskala besar ini menggunakan DL untuk mengidentifikasi limfosit di seluruh gambar dan menghubungkan karakteristik spasial limfosit dengan penilaian molekuler. Artikel ini adalah kunci untuk sistem otomatis kuantifikasi sel imun dari slide H&E dan identifikasi sub-kategori infiltrasi imun yang terkait dengan hasil terapi.
102. Corredor, G. dkk. Arsitektur spasial dan pengaturan limfosit yang menginfiltrasi tumor untuk memprediksi kemungkinan kekambuhan pada kanker paru non-sel kecil stadium awal. *Clin. Cancer Res.* **25**, 1526-1534 (2018).
Dalam makalah ini, pengaturan spasial, dan bukan hanya kepadatan, limfosit yang menyusup ke dalam tumor pada gambar patologi kanker paru stadium awal terbukti dapat menjadi prognostik kekambuhan. Perbandingan yang komprehensif disediakan, menunjukkan bahwa fitur yang diekstraksi komputer dari pengaturan spasial limfosit yang menyusup ke dalam tumor lebih prognostik
daripada penghitungan manual (ahli patologi) dari kepadatan limfosit yang menyusup ke dalam **tumor.**
103. Cohen, O., Zhu, B. & Rosen, M. S. MR sidik jari Deep RecOnstruction NEtwork (DRONE). *Magn. Reson. Med.* **80**, 885-894 (2018).
104. Chen, H. dkk. CT dosis rendah dengan jaringan saraf konvolusi encoder-decoder residual (RED-CNN). Tersedia di *arXiv* https://arxiv.org/abs/1702.00288 (2017).
105. Coudray, N. dkk. Klasifikasi dan prediksi mutasi dari citra histopatologi kanker paru non-sel kecil menggunakan deep learning. *Nat. Med.* **24**, 1559-1567 (2018).
Makalah ini menggunakan kerangka kerja DL untuk memprediksi mutasi dari gambar H&E, yang berimplikasi untuk mengidentifikasi wawasan mekanistik utama dari pencitraan seluruh slide standar serta untuk stratifikasi pasien.
106. Turkki, R., Linder, N., Kovanen, PE, Pellinen, T. & Lundin, J. Pembelajaran mendalam yang diawasi oleh antibodi untuk kuantifikasi sel imun yang menginfiltrasi tumor dalam sampel kanker payudara yang diwarnai dengan hematoksilin dan eosin.
J. Pathol. Inform. **7**, 38 (2016).
107. Norgeot, B., Glicksberg, BS & Butte, AJ Panggilan untuk perawatan kesehatan dengan pembelajaran mendalam. *Nat. Med.* **25**, 14-15 (2019).
108. Esteva, A. dkk. Panduan untuk pembelajaran mendalam dalam perawatan kesehatan.
Nat. Med. **25**, 24-29 (2019).
109. Yang, Z. dkk. Diagnosis asisten klinis untuk rekam medis elektronik berdasarkan jaringan syaraf tiruan convolutional . *Sci. Rep.* **8**, 6329 (2018).
110. Steele, AJ, Denaxas, SC, Shah, AD, Hemingway, H. & Luscombe, NM Model pembelajaran mesin dalam catatan kesehatan elektronik dapat mengungguli model kelangsungan hidup konvensional untuk memprediksi mortalitas pasien pada penyakit arteri koroner. *PLOS ONE* **13**, e0202344 (2018).
111. Mohr, D. C., Zhang, M. & Schueller, S. M. Penginderaan pribadi: memahami kesehatan mental menggunakan sensor yang ada di mana-mana dan pembelajaran mesin. *Annu. Rev Clin. Psychol.* **13**, 23-47 (2017).
112. Gkotsis, G. dkk. Karakterisasi kondisi kesehatan mental di media sosial menggunakan Informed Deep Learning. *Sci. Rep.* **7**, 45141 (2017).
113. Koscielny, S. Mengapa sebagian besar tanda tangan ekspresi gen tumor belum berguna di klinik. *Sci. Transl Med.* **2**, 14ps12 (2010).
114. Odell, SG, Lazo, GR, Woodhouse, MR, Hane, DL & Sen, TZ Seni kurasi pada basis data biologi: prinsip dan aplikasi. *Curr. Plant Biol.* **11-12**, 2-11 (2017).

Ucapan Terima Kasih
Para penulis berterima kasih kepada E. Birney dan E. Papa atas komentar-komentar yang sangat membantu, M. Segler yang telah berkontribusi pada subbagian opti-misasi molekul kecil dan A. Janowczyk yang telah menyediakan gambar-gambar patologi pada Gambar 4.

Kepentingan yang bersaing
Para penulis menyatakan tidak memiliki kepentingan yang bersaing.

Catatan penerbit
Springer Nature tetap netral sehubungan dengan klaim yurisdiksi dalam peta yang dipublikasikan dan afiliasi kelembagaan.

TAUTAN TERKAIT
DeepChem: https://www.deepchem.io/
Tantangan DREAM: http://dreamchallenges.org/
TensorFlow: https://www.tensorflow.org/